

e-ISSN: 2148-7456

a peer-reviewed  
online journal

hosted by DergiPark

# International Journal of Assessment Tools in Education

**Volume: 12**

**Issue: 4**

**December 2025**

<https://dergipark.org.tr/en/pub/ijate>



e-ISSN: 2148-7456

**Volume 12**

**Issue 4**

**December 2025**

**Editor** Dr. Safiye BILICAN DEMIR  
**Address** Department of Educational Sciences, Kırıkkale University, Ankara yolu Km 7. Yahşihan 71450, Kirikkale, Türkiye

**E-mail** [ijate.editor@gmail.com](mailto:ijate.editor@gmail.com)  
[safiyebilican@gmail.com](mailto:safiyebilican@gmail.com)

**Publisher** Dr. Izzet KARA  
**Address** Pamukkale University, Education Faculty, Kinikli Campus, 20070, Denizli, Türkiye

**Phone** +90 258 296 1036  
**Fax** +90 258 296 1200  
**E-mail** [ikara@pau.edu.tr](mailto:ikara@pau.edu.tr)  
[ijate.info@gmail.com](mailto:ijate.info@gmail.com)

**Journal Contact** Dr. Eren Can AYBEK  
**Address** Department of Educational Sciences, Pamukkale University, Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye  
**Phone** +90 258 296 31050  
**Fax** +90 258 296 1200  
**E-mail** [erencanaybek@gmail.com](mailto:erencanaybek@gmail.com)

**Address** Dr. Anil KANDEMİR  
Department of Educational Sciences, Agri Ibrahim Cecen University, Faculty of Education, Agri, Türkiye  
[akandemir@agri.edu.tr](mailto:akandemir@agri.edu.tr)

**Frequency** 4 issues per year (March, June, September, December)

**Online ISSN** 2148-7456

**Website** <https://dergipark.org.tr/en/pub/ijate>  
<https://ijate.net/index.php/ijate>

**Cover Design** IJATE

International Journal of Assessment Tools in Education (IJATE) is a peer-reviewed and academic online journal. The scientific and legal responsibility for manuscripts published in our journal belongs to the authors(s).

## International Journal of Assessment Tools in Education

*International Journal of Assessment Tools in Education* (IJATE) accepts original research on the design, analysis and use of evaluation along with assessment to enhance comprehension of the performance and quality of stakeholders in educational settings. IJATE is pleased to receive discriminating theoretical and empirical manuscripts (quantitative or qualitative) which could direct significant national and international argumentations in educational policy and practice.

IJATE, as an online journal, is hosted by DergiPark [TUBITAK-ULAKBIM (The Scientific and Technological Research Council of Türkiye)].

In IJATE, there are no charges under any procedure for submitting or publishing an article.

### Indexes and Platforms:

- Emerging Sources Citation Index (ESCI)
- TR Index (ULAKBIM),
- Education Resources Information Center (ERIC),
- EBSCOhost,
- SOBIAD Citation Index,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,

## Editor

Dr. Safiye BILICAN DEMIR, *Kırıkkale University, Türkiye*

## Section Editors

Dr. Ebru BALTA, *Ağrı İbrahim Çeçen Üniversitesi, Türkiye*

Dr. Nisu BORAN, *Pamukkale University, Türkiye*

Dr. Sumeyra SOYSAL, *Necmettin Erbakan University, Türkiye*

Dr. Ömer Faruk ŞEN, *Kırıkkale University, Türkiye*

Dr. Selma SENEL, *Balikesir University, Türkiye*

Dr. Esin YILMAZ KOGAR, *Nigde Omer Halisdemir University, Türkiye*

## Editorial Board

Dr. Beyza AKSU DUNYA, *Bartın University, Türkiye*

Dr. Stanislav AVSEC, *University of Ljubljana, Slovenia*

Dr. Utkun AYDIN, *University of Glasgow, United Kingdom*

Dr. Duru BAYRAM, *Eindhoven University of Technology, The Netherlands*

Dr. Kelly D. BRADLEY, *University of Kentucky, United States*

Dr. Okan BULUT, *University of Alberta, Canada*

Dr. William W. COBERN, *Western Michigan University, United States*

Dr. Nuri DOGAN, *Hacettepe University, Türkiye*

Dr. Selahattin GELBAL, *Hacettepe University, Türkiye*

Dr. Qiwei HE, *Georgetown University, United States*

Dr. Lim HOOI LIAN, *School of Educational Studies University Sains, Malaysia*

Dr. Tugba KARADAVUT, *Izmir Democracy University, Türkiye*

Dr. Orhan KARAMUSTAFAOGLU, *Amasya University, Türkiye*

Dr. Hulya KELECIOGLU, *Hacettepe University, Türkiye*

Dr. Hakan KOGAR, *Akdeniz University, Türkiye*

Dr. Seongyong LEE, *Hannam University, South Korea*

Dr. Sunbok LEE, *University of Houston, United States*

Dr. Hamzeh MORADI, *Sun Yat-sen University, China*

Dr. Nesrin OZTURK, *Izmir Democracy University, Türkiye*

Dr. Turan PAKER, *Pamukkale University, Türkiye*

Dr. Hossein SALARIAN, *University of Tehran, Iran*

Dr. Saranya T.S., *Amity University Bangalore, India*

Dr. Ragip TERZI, *Harran University, Türkiye*

Dr. Turgut TURKDOGAN, *Pamukkale University, Türkiye*

Dr. Hongwei YANG, *University of West Florida, United States*

Dr. Ozen YILDIRIM, *Pamukkale University, Türkiye*

## English Language Editors

Dr. Hatice ALTUN, *Pamukkale University, Türkiye*

Ahmet KUTUK, *Akdeniz University, Türkiye*



**Editorial Assistant**

Dr. Asiye BAHTIYAR, *Pamukkale University, Türkiye*

Dr. Neslihan Tugce OZYETER, *Kocaeli University, Türkiye*

PhDc. Ibrahim Hakki TEZCI, *Akdeniz University, Türkiye*

**Technical Assistant**

Dr. Eren Can AYBEK, *Pamukkale University, Türkiye*

Dr. Anil KANDEMİR, *Agri Ibrahim Cecen University, Türkiye*

## CONTENTS

### Research Articles

1. Effectiveness of a parenting program on children's academic self-control and parents' competence

**Page: 871-890 PDF**

Hanife Büyük, Selahiddin Öğülmüş

2. Mathematical literacy, thinking, and teacher self-efficacy: A latent profile analysis of preservice teachers' competencies

**Page: 891-904 PDF**

Muhammed Celal Uras, Mehmet Şata, Yasin Soylu

3. Answer-based and reference-based BERT models for automatic scoring of Turkish short answers: The decisive role of task complexity

**Page: 905-925 PDF**

Abdulkadir Kara, Zeynep Avinç Kara, Serkan Yıldırım

4. From addiction to pervasiveness: Validation of the Smartphone Pervasiveness Scale in Turkish adolescents

**Page: 926-941 PDF**

Osman Urfa, Recep Görgülü

5. Exploring trends in psychometrics literature through a structural topic model

**Page: 942-962 PDF**

Kübra Atalay Kabasakal, Rabia Akcan, Duygu Koçak

6. Support for Gender Equality among Men Scale: Adaptation to Turkish culture

**Page: 963-982 PDF**

Esmâ Esen Çiftçi, Esra Daşcı, Cansu Ayan, Zeynep Uludağ

7. Scientific article review platform using generative artificial intelligence to streamline the peer review process

**Page: 983-994 PDF**

German Cuaya-simbro, Serguei Drago Domínguez Ruiz

8. Developing Turkish-Sign-Language Self-Efficacy Scale for Learners based on various test theories

**Page: 995-1015 PDF**

Pelin Piştav Akmeşe, Asiye Şengül Avşar, Nilay Kayhan, Necla Işıkdöğün Uğurlu, Zeynep Oral

9. The development of the Sustainable Consumption Behavior and Intention Scale

**Page: 1016-1033 PDF**

Merve Eker Çelebi, Fatma Taşkın Ekici

10. Use of ASSURE MODEL in ELT: Reflections on the learning and teaching process

**Page: 1034-1054 PDF**

Hatun Vera Akşab, Melike Özyurt

11. ChatGPT vs. DeepSeek: A comparative psychometric evaluation of AI tools in generating multiple-choice questions

**Page: 1055-1079 PDF**

Ceylan Gündeğer Kılıcı

12. Construction and validation of a multilingual diagnostic instrument for neuromyths and their origins

**Page: 1080-1105 PDF**

Oktay Cem Adıgüzel, Patrice Potvin, Sibel Küçükkayhan, Derya Atik Kara

13. Deficit Thinking Scale for teachers: A validity and reliability study in Turkish context

**Page: 1128-1147 PDF**

Abide Ocak, İsmail Çimen

14. The effect of STEM practices on students' attitudes and achievements: A meta-analysis study

**Page: 1148-1169 PDF**

Abdulkadir Kurt, Muhammed Akıncı

15. Investigating homogeneity of variance in normal, skewed-normal, and gamma distributions: A simulation study

**Page: 1170-1185 PDF**

Serpil Çelikten Demirel, Ayşenur Erdemir, Esra Oyar, Tuba Gündüz

**Review Articles**

1. Video annotation tools for assessing psychomotor skills in nursing education: A scoping review

**Page: 1106-1127 PDF**

Greet Leysens, Rani Claus, Wim Van Petegem, Nathalie Charlier

## Effectiveness of a parenting program on children's academic self-control and parents' competence

Hanife Büyük<sup>1\*</sup>, Selahiddin Öğülmüş<sup>2</sup>

<sup>1</sup>Ministry of National Education, Ankara, Türkiye

<sup>2</sup>Ankara University, Faculty of Educational Sciences, Division of Educational Psychology, Ankara, Türkiye

### ARTICLE HISTORY

Received: Aug. 5, 2024

Accepted: July 5, 2025

### Keywords:

Academic self-control,  
Parent training,  
Family academic support,  
Middle school students,  
Academic success.

**Abstract:** Self-control is a determining factor for academic success. Fortunately, it can be strengthened through positive parenting practices, especially during childhood. This study aims to investigate the effects of a parent training program designed to enhance the academic self-control skills of children aged 11 to 14. The research had an experimental design with pre-test and post-test for experimental ( $n = 11$ ) and control groups ( $n = 11$ ) attended randomly. The program was held online, individually, and once a week for eight weeks. A three-month follow-up data was collected from the experimental group. In this study, Academic Self-Control Scale and Parenting Sense of Competence Scale were used. In analysis, the difference scores of the experimental and control groups were compared using Mann-Whitney U Test; for the difference between the experimental group pre-test, post-test and three-month follow-up mean scores the Friedman Test were run. According to the findings, it was observed that the academic self-control total scores and academic perseverance sub-scores of the children and parenting competence scores of the mothers in the experimental group increased significantly after the intervention compared to the control group.

## 1. INTRODUCTION

In today's educational landscape, students spend extended hours at school and are expected to meet a wide range of academic demands, including attending classes, reviewing course material, submitting assignments, preparing projects, and taking examinations. The competitive structure of the education system requires substantial effort to meet established academic goals and succeed. This pursuit of academic achievement, which begins in the early years of formal education, continues throughout a student's academic trajectory.

Achieving success involves more than cognitive ability; it also necessitates the use of effective study strategies and sustained effort. Academic success is a critical factor that shapes access to higher education, influences career opportunities, and affects various aspects of personal and professional life. It may even impact the educational prospects of future generations (Feinstein *et al.*, 2008). As a result, identifying the variables that influence academic success has become a central concern in educational research.

---

\*CONTACT: Hanife BÜYÜK ✉ [hanifeyasav@gmail.com](mailto:hanifeyasav@gmail.com) 📍 Ministry of National Education, Ankara, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

Self-control is recognized as a key predictor of academic success (Duckworth & Seligman, 2005; Robson *et al.*, 2020; Vazsonyi *et al.*, 2021). Muraven and Baumeister (2000) define self-control as the capacity to deliberately alter one's emotions, thoughts, and behaviors in pursuit of long-term goals. When a discrepancy arises between one's current state and desired outcomes, individuals are motivated to adjust their internal states and external actions. In such cases, self-control entails the conscious effort required to resolve this conflict through behavioral regulation (Baumeister & Vohs, 2007).

Self-control influences not only academic performance but also various other domains of life. It has been linked to consistent work habits, adherence to dietary routines, the maintenance of healthy interpersonal relationships (Tangney *et al.*, 2004), organizational skills, and problematic behaviors such as obesity, alcohol use, gambling, and drug addiction (Galla & Duckworth, 2015). Core abilities such as achieving goals, planning, impulse regulation, behavioral inhibition, and managing social conduct are strongly associated with self-control. Individuals with higher levels of self-control are more capable of resisting distractions, sustaining focus, and disengaging from obsessive or intrusive thoughts (Smithers *et al.*, 2018; Steinberg, 2018). Moreover, self-control shows positive associations with academic competence, psychological adjustment, and the avoidance of substance use and risky sexual behavior (Dishion & McMahon, 1998).

### 1.1. Academic Self-Control

Students who are able to focus on lessons, ignore distractions while studying, complete their assignments regularly, prepare for exams more meticulously, work in a planned manner, and set academic goals are more likely to achieve higher grades compared to their peers (Duckworth *et al.*, 2019; Tangney *et al.*, 2004). The ability to postpone that desire (e.g. putting the phone away because of an exam the next day) and to wait with an awareness of the long-term gain (e.g. getting a good grade on the exam), rather than experiencing the short-term pleasure of the current desired action (e.g. checking social media) contributes significantly to personal success. Self-control is a skill that requires students to complete their academic tasks without procrastination and to set aside impulsive desires that offer momentary gratification. The relationship between self-control and academic success, and its comparison with intelligence in this context, highlights the importance of self-control in academic settings. The self-control behavior of individuals in the academic context is referred to as "academic self-control" (Rebecca, 2015).

Students with high academic self-control tend to set personal goals, maintain a positive outlook when facing difficulties and expectations, and trust both social and nonsocial sources of support (e.g., going to the library). They take notes, organize their environment to promote learning, rely on self-evaluation strategies (e.g., rewarding themselves), and review written materials (Galla *et al.*, 2018). Even when they perform poorly on an exam or assignment, they are likely to reflect on the possible causes of failure and adjust their learning goals and strategies accordingly (Kennett & Maki, 2014).

According to Kennett's (1994) Academic Self-Control Model, individual self-control and self-efficacy, defined as the regulation of one's cognitions, meet the academic challenges and obstacles faced by students. These two systems interact and jointly give rise to both self-control behaviors and academic hardship behaviors. In turn, these behaviors contribute to positive academic outcomes. According to the model, the components of academic self-control are academic strength, general resourcefulness, academic self-efficacy, and attributions for failure (lack of effort, difficulty of task, bad luck/fate, lack of talent). On the other hand, it can be said that although all of these variables are necessary and interact, general resourcefulness skills are fundamental in explaining academic self-control (Kennett, 1994; Rosenbaum, 1990, 2000). In short, no matter how strongly a student believes in their academic abilities (self-efficacy) or attributes past failures to effort and skill (attribution), the ability to change habits ultimately

depends on personal initiative. Without essential skills, such as delaying immediate gratification or engaging in positive self-talk, it becomes difficult to manage everyday life challenges and academic demands (Rebecca & Kennett, 2018).

Responding sensitively to the child's needs and signals is associated with various developmental outcomes, ranging from secure attachment and social-emotional growth to cognitive development and academic success (Sroufe, 2005). Through appropriate behaviors and practices, parents can help their children learn to regulate their emotions, thoughts, and behaviors. High levels of parental acceptance, consistent behavioral control, and parental monitoring, as well as minimal use of psychological control, create a supportive environment for fostering a child's self-control. Family-related factors, including parent-adolescent relationships, adequate parental monitoring, and positive parenting practices play a central role in preventing negative developmental outcomes in adolescence (Cho *et al.*, 2018). Parental practices play a crucial role in the development of self-control. Parents who express love and warmth, offer support, show approval and acceptance, and consistently enforce rules and boundaries create the conditions for children to learn how to follow rules effectively and delay immediate gratification (Finkenauer *et al.*, 2010).

Self-control skills can change over time, influenced by familial and external factors. While Gottfredson and Hirschi (1990) emphasize early socialization in the development of self-control, they also acknowledge that self-control is a form of individual difference or biological variability. There is also a strong genetic basis in the persistence of self-control over time (Coyne & Wright, 2014). The strong influence of genetics does not suggest that socialization processes, including effective parenting, are insignificant in the development of self-control. Moreover, biological factors are known to predict developmental outcomes in interaction with environmental influences (Vazsonyi & Huang, 2010).

Several intervention programs aimed at enhancing self-control incorporate specific self-regulation strategies. These strategies, both cognitive and situational, are based on Bandura's Social Cognitive Theory (1977). The primary objective of self-control techniques, such as behavioral self-control training, cognitive self-regulation, and self-management techniques, is to reduce behavioral deficits or excesses. In a study examining the impact of using self-regulation strategies on academic goal achievement (Williamson, 2021), the most commonly used situational strategies were distraction elimination and distraction avoidance. The most frequently employed cognitive strategy was goal focusing. In the process of developing self-control, it is recommended that rewards be associated with effort, enabling individuals to gradually derive satisfaction from the effort itself. According to Strayhorn (2002), individuals can use techniques such as imaginary rehearsal, learning adaptive skills, and introspection to avoid being misled by momentary physical or mental distractions and to regulate their behavior.

There are several studies focusing on self-control development in school-aged children through parental and teacher involvement, particularly by setting limits and assigning responsibilities (Duckworth *et al.*, 2014). It is argued that the most critical period for learning self-control and problem-solving strategies is middle childhood (Piquero *et al.*, 2016). Interventions implemented during this stage are considered especially effective in producing lasting behavioral changes (Moffitt *et al.*, 2011), as middle childhood coincides with the pre-adolescent developmental phase.

In order to protect children from risks and ensure that they benefit fully from opportunities, it is necessary to address not only the children themselves but also the environments in which they develop. Adolescence is an important period for risk and protective factors (O'Connell *et al.*, 2009; Zahn-Waxler *et al.*, 2000). The fact that early adolescence coincides with the transition period from elementary to secondary school can be challenging as children encounter new academic demands. During this period, parents' interactions, emotional validation, and appropriate guidance are essential. Such guidance should aim to strike a balance between what

the child is currently capable of and what they can potentially accomplish. This supportive process, known as "scaffolding" (Vygotsky, 1984), involves the gradual transfer of autonomy and responsibility to the child. As the children's self-regulation abilities develop, the degree of parental control should correspondingly decrease and eventually be withdrawn. Effective parental scaffolding fosters the development of self-control by shaping children's executive functioning and emotion regulation skills during critical developmental periods. This theoretical perspective highlights that while self-control may have a genetic basis, its development and long-term stability are profoundly shaped by early socialization experiences and the quality of caregiver-child interactions (Moffitt *et al.*, 2011).

Therefore, it is important to foster adolescents' capacity for self-control, enabling them to experience a more successful and healthy developmental trajectory in both psychosocial and educational domains. Studies indicate that self-control continues to develop throughout adolescence as a function of the adolescent's neuroplasticity, that parents play an important role in this development, and that positive parenting behaviors contribute to the enhancement of self-control (Beaver *et al.*, 2010; Bolger *et al.*, 2022). Parental academic support closely reflects parents' expectations for their children and the value they place on academic success, and it also represents a culturally embedded approach (Chen, 2005). In a collectivist cultural orientation, families tend to interpret a child's academic success or failure as reflective of the entire family (Kağıtçıbaşı, 2005); consequently, their willingness to support the child is typically high. It has been found that when parents create a suitable environment at home for studying (Gottfried *et al.*, 1994) and when they are more involved in their children's school-related activities (Chen, 2005), children show greater academic achievement and school adjustment.

In Türkiye, an educational program including positive parental behaviors to support directly academic self-control of children wasn't come across. This study was carried out to contribute to the elimination of this gap in self-control literature. The problem of this study is the lack of a psychoeducation program for parents that has been developed and tested as effective to support students' academic self-control skills through their parents. To this end, it has been investigated whether the Family Support on Academic Self-Control Program (FSASC) prepared for parents of middle school students aged between 11-14, is effective in developing the academic self-control skills of the children. In this connection, answers to the following questions were sought:

1. Do the academic self-control skills of the children whose parents participated in the training improve significantly after the training?
2. Do the parenting competence of the parents who participated in the training improve significantly after the training?

## 2. METHOD

### 2.1. Research Model

This study is an experimental research conducted within the quantitative research methodology. Pre-test and post-test measurements were taken with both experimental and control groups. Follow-up measurements of the experimental group were taken three months after the post-tests. In experimental designs, cause-and-effect relationships between variables are examined. In such studies, it is desirable that the groups are equivalent in terms of participant characteristics and external variables. If it is ensured that the groups are equivalent in all aspects except the independent variable applied, establishing causality is more meaningful. The most important and best control technique is accepted as random assignment (Christensen *et al.*, 2015). In the current study, participants were randomly assigned to the groups.



## 2.2. Procedure

This study aimed to support children's academic self-control skills through their parents. For this purpose, a psychoeducational program designed to enhance positive parenting skills was developed and implemented for parents. Due to the conditions imposed by the COVID-19 pandemic, the training was delivered individually and online. Based on expert opinions, certain group activities originally designed for face-to-face sessions were adapted for online and individual implementation. PowerPoint presentations were prepared for the program and shared with participants via the Zoom platform. Daily sharings and contact between the trainer and the participants was maintained using WhatsApp and e-mail. A total of 22 mothers participated in the program (11 in the experimental group and 11 in the control group).

Although the training was provided exclusively to parents, data were collected from both children and their parents. The study examined whether there were changes in children's academic self-control skills and parents' parenting competence levels before and after the training. Follow-up measurements were conducted three months after the training. The research design is shown in Table 1. The control group also received the training after the post-test assessments were completed.

**Table 1.** Research design.

Group	Assignment	Measurement	Treatment	Measurement	Measurement
Experimental group	R	Pre-test	FSASC	Post-test	Follow-up
Control group	R	Pre-test	-	Post-test	-

R: Random assignment

## 2.3. Study Group

The study was conducted through a random sampling method. Initial contact was made with volunteer families from three different state middle schools in Etimesgut district of Ankara province via email and WhatsApp using a poster of the program. Since the age group was limited 11–14, 5th and 8th grade students were excluded. Contact was made only with the parents of 6th and 7th grade students. Fifth graders may be younger than 11 years old so they were excluded due to age. Meanwhile, 8th grade students are dealing with the high school entrance exam in Türkiye and their families may be busy and focused on it. Therefore, they were also excluded from the study.

Initially, only 29 mothers expressed interest in participating in the program, while the fathers were reluctant. Five mothers who did not want to complete the pre-test measurements for various reasons were excluded. As a result, 24 mothers (12 in the experimental group and 12 in the control group) were randomly assigned to the groups. One participant from the experimental group did not continue the training after the first session, so the program was completed with 11 mothers using the Zoom platform with one session per week for eight weeks. Following the training, post-tests were administered to 11 mothers from the experimental group and 11 mothers from the control group (excluding one mother who did not complete the post-tests).

After the post-test measurements, for study ethical reasons, the training was offered to seven mothers from the control group who wished to participate. In summary, the study group consists of mothers ( $n = 22$ ; 11 in experimental group, 11 in control group) and their children aged 11–14 ( $n = 22$ ; 11 in experimental group, 11 in control group). The average age of the mothers is 39 years, and of children is 11.73 years. In the experimental group, 63.64% ( $n = 7$ ) of the children are girls, while in the control group, 54.55% ( $n = 6$ ) are girls. Half of the children in the study group are in the 6th grade, the other half are in the 7th grade.

## 2.4. Data Collection Tools

Data were collected from both parents and children within the study. Rather than demographic information, the level of parenting skills of parents was measured using Parenting Sense of



Competence Scale (Gibaud-Waliston & Wandersman, 1978), and the academic self-control level of children was evaluated with Academic Self-Control Scale (ASCS), which was developed by the researchers (Büyük *et al.*, 2020). Additionally, a psychoeducational program “Family Support on Academic Self-Control Program” (FSASC), was developed to improve children's academic self-control through parental support.

#### 2.4.1. Parenting Sense of Competence Scale

The Parenting Sense of Competence Scale (PSCS) developed by Gibaud-Waliston and Wandersman (1978) was used to assess parents' satisfaction with their parenting and their self-efficacy in the parenting role. The PSCS consists of two subscales- satisfaction and effectiveness/usefulness- and contains 16 items like “Being a parent is manageable, and any problems are easily solved.” Items are rated on a 6-point Likert scale ranging from 1 (“*Strongly Disagree*”) to 6 (“*Strongly Agree*”). On the PSCS, items 2, 3, 4, 5, 8, 9, 12, 14, and 16 are reverse coded.

The adaptation studies for Turkish culture were conducted by Çokamay and Kapçı (2016). The Turkish version of the scale was administered to a sample of 383 parents as part of the adaptation study. Analyses revealed that the scale preserved its original structure and consisted of two subdimensions. The Cronbach's alpha internal consistency coefficient for the Turkish version was calculated as .75 for the entire scale. To assess test-retest reliability, the scale was administered to 46 parents at a three-week interval. Pearson correlation analysis indicated a reliability coefficient of  $r = .55$  for the total score ( $p < .01$ ).

An increase in the score obtained from the scale indicates an increase in parenting competence. In this study, the total score of the PSCS was used.

#### 2.4.2. Academic Self-Control Scale

The Academic Self-Control Scale (ASCS) was developed to evaluate the academic self-control behaviors of middle school students between the ages of 11 and 14 (Büyük *et al.*, 2020). The ASCS consists of 12 items across two subscales, using a 5-point Likert-type response format ranging from “never” to “always.” The two subscales are labeled “academic perseverance” and “academic attention.” These factors together explained 47.11% of the variance in participants' academic self-control scores.

Structural validity of the scale was supported by Principal Component Analysis (PCA). In the initial analysis, the Kaiser–Meyer–Olkin (KMO) measure confirmed sampling adequacy (.86), and Bartlett's test of sphericity was significant,  $\chi^2(210) = 1258.74$ ,  $p < .001$ . Factor loadings ranged from .566 to .806 for academic perseverance and from .620 to .743 for academic attention.

Confirmatory Factor Analysis (CFA) confirmed the suitability of the factor structure. The CFA was conducted on a separate sample ( $n = 297$ ;  $M = 12.92$ ,  $SD = 0.97$ ), and model fit indices indicated an acceptable fit:  $\chi^2/df = 2.56$ , RMSEA = .065, CFI = .93, TLI = .91, and SRMR = .05. All standardized factor loadings were statistically significant (t-values ranging from 7.94 to 11.65). Average Variance Extracted (AVE) values supported convergent validity: .83 for academic perseverance and .80 for academic attention.

The Academic Perseverance Scale (Bozgün & Başgöl, 2018) was used to assess criterion validity. A strong positive correlation was found between the total scores of the ASCS and the Academic Perseverance Scale ( $r = .74$ ,  $p < .001$ ). Subscale correlations also revealed a strong association between academic perseverance scores from both scales ( $r = .71$ ,  $p < .001$ ) and a moderate correlation with academic attention ( $r = .46$ ,  $p < .001$ ).

The internal consistency reliability, measured using the Cronbach alpha coefficient, was found to be .81 for the entire scale, .81 for the academic perseverance factor, and .64 for the academic attention factor. Test–retest reliability was evaluated using a four-week interval with 54

students ( $n = 54$ ;  $M = 12.52$ ,  $SD = 0.90$ ). The resulting coefficients were .93 for the total scale, .89 for academic perseverance, and .91 for academic attention. Composite reliability values were also acceptable: .80 for perseverance and .70 for attention.

The scale includes reverse-coded items, and the total score ranges from 12 to 60. Higher scores reflect greater levels of academic self-control. An example item is: “*My friends say I act out in a disturbing way during class.*”

#### **2.4.3. Family support on academic self-control program**

Within the study, a psychoeducational program titled “Family Support on Academic Self-Control” (FSASC) was developed by the researchers to improve children's academic self-control through parental support. The development process of the FSASC lasted longer than usual and was challenging due to the COVID-19 pandemic. Initially it was developed as a face-to-face group program. However, due to difficulties in finding participants, it was first adapted into an online group program, and eventually implemented in that format.

Before the development of FSASC literature reviews, observations, interviews, focus group studies and expert opinions were evaluated. The researcher's own observations, along with individual and group interviews with middle school students and parents, helped determine the objectives of the program. Expert consultations were obtained to ensure content validity. Both national and international programs on education and self-control were reviewed, with a specific focus on interventions designed for adolescents.

Based on expert feedback, some group activities originally intended for face-to-face settings were adapted for online individual sessions. PowerPoint presentations were prepared and shared with participants via the Zoom platform. Daily communication between the trainer and participants was maintained through WhatsApp and e-mail.

The theoretical foundation of the program is based on Cognitive Behavioral Therapy (CBT), with strong influences from social learning theories and behaviorist principles. Sessions on active listening and emotional acceptance also incorporate mindfulness techniques, as proposed by Jon Kabat-Zinn (2003), aiming to enhance self-awareness for both parents and children.

The overall goal of the program is to strengthen children's academic self-control through their parents, focusing on the development of positive parenting strategies. Participants are supported in becoming parents who are aware of their child's developmental needs, understand the consequences of their behavior toward their child, and recognize their responsibility in fostering skills essential for academic success. To achieve these goals, the program includes practical recommendations that can be applied in daily life. Great care was taken to ensure the program's feasibility and its relevance to everyday parenting contexts.

FSASC consists of eight sessions, including an introductory and a concluding session. Each session, except the introduction, lasts 120 minutes. Sessions are held once a week at times suitable for the trainer and participants—typically evenings after 9 p.m. or weekends for working mothers, and daytime on weekdays for non-working mothers.

The program begins with an introductory session in which the purpose, structure, content, participant responsibilities, and participation rules are discussed. The session concludes with participants sharing their expectations for the program and identifying the parenting behaviors they would like to change or maintain in relation to their children.

Following the introduction, the *first session* focuses on *growing up and self-control*. It begins with the “Good Teacher – Bad Teacher” activity, in which participants describe behaviors of both good and bad teachers. This exercise helps them draw parallels between parenting and teaching, distinguish between effective and ineffective approaches, and reflect on their own parenting behaviors. The session also introduces key developmental changes during adolescence and includes the activity “My Child is Changing” to raise awareness. The concept of self-control is defined with everyday examples, and participants share personal experiences

involving self-control. The session concludes with an action plan: participants are asked to observe and record their own "good teacher" and "bad teacher" behaviors toward their child over the following week.

The *second session* starts with a discussion of the first action plan. The theme is *positive parenting*, defined through warmth and sensitivity. Participants are encouraged to evaluate their own parenting behavior and analyze their child's behavior using behavioral analysis. They practice identifying preconditions that lead to negative behavior using example scenarios, and they discuss real-life cases involving their children. Together, they explore faulty parenting strategies and appropriate consequences from both child and parent perspectives. The session includes the "Emotional Tide" activity to deepen understanding. It ends with participants completing a behavior analysis as a second action plan.

The *third session* addresses *setting boundaries*. Participants learn the importance of boundaries in academic life and practice appropriate methods for establishing them. Through the "3D Form" (Emotion–Behavior–Thought), they analyze the links between emotional states and actions. They also role-play emotional validation and content validation using structured examples. The trainer facilitates a three-stage boundary-setting exercise, followed by role-playing scenarios specific to school-related situations. The action plan focuses on setting academic self-control boundaries with their children during the week.

The *fourth and fifth sessions* emphasize *strengthening behavioral self-regulation*. In the fourth session, participants reflect on how they and their children manage self-control in four life areas: diet, exercise, media use, and sleep. They identify weaker areas and learn how to reinforce them. Topics such as modeling healthy behaviors and managing internet use, nutrition, and sleep habits are discussed. Participants are given specific action plans for setting limits in each of the four domains.

The *fifth session* focuses on *organization and consistency*. Participants explore how routines and rules support self-control. They are encouraged to involve children in household tasks to build responsibility, and they are provided with a sample task list. A self-control agreement is drafted, reinforcing rule-following and duty completion through role play. As an action plan, participants create family rules and agreements to be implemented at home.

The *sixth session* centers on *academic self-control*. Topics include goal setting, planning, self-monitoring, self-reinforcement, and self-evaluation. The "Whose Responsibility Is It?" chart helps clarify parent and child roles in education. Participants explore how to regulate the home environment to promote study routines and complete the "My Goals" activity. They review examples of a self-control diary and lesson plan. The action plan includes working with their child to develop academic goals, create a study schedule, and maintain a self-monitoring diary.

The *final session* includes a group reflection on the previous week's activities and a full evaluation of the program. Participants discuss the sessions and activities they found most helpful or challenging. They are encouraged to continue practicing what they learned. At the conclusion, they receive a certificate of participation prepared by the researcher.

## 2.5. Data Analysis

In the current study, the quantitative analysis method was used. In this context, the difference between the pre-test and post-test mean scores of the experimental and control groups was analyzed. In addition, it was analyzed whether there was a significant difference between the experimental group pre-test, post-test, and three-month follow-up mean scores. Prior to analysis, the dataset was examined for missing values and assumptions of normality. No missing data were identified. Normality was assessed using skewness and kurtosis coefficients (acceptable range:  $-1$  to  $+1$ ) and the Shapiro–Wilk test. The results indicated that while some variables met the assumption of normality, others did not.

Given the small sample size ( $n = 22$ ), non-parametric tests were preferred in line with Büyükoztürk (2009), who recommends parametric analysis when each group includes at least 15 participants. The Mann–Whitney U test was used to compare the pre- to post-test change scores between the experimental and control groups. The Wilcoxon Signed Ranks test was used to test the results of the pre-test, and post-test scores of the experimental and control groups separately. The relationships between two independent numerical variables were analyzed using Spearman's Rho correlation coefficient. The average scores of the experimental group pre-test, post-test and follow-up were analyzed using the Friedman Test. In the case of a significant difference in the Friedman Test, post-hoc analyses were performed to determine which measurement caused the difference, and Bonferroni-corrected results were used. If the difference between the means was found to be significant after each analysis, an effect size calculation was also performed. The effect size calculation proposed by Field (2017) for the Mann Whitney U Test and the Wilcoxon Signed Ranks Test was performed using the formula  $r = z / \sqrt{N}$ . The following formula was used to calculate the effect size according to the Friedman analysis:  $W = \chi^2 / n (K-1)$ . As stated by Cohen (1988) for the  $r$  and  $W$  values obtained as a result of the calculation: .1 = small effect, .3 = medium effect, .5 = large effect. The SPSS 22.00 software was used to analyze the data, and statistical significance was assumed to be  $p < .05$ .

### 3. RESULTS

#### 3.1. Findings Related to the Academic Self-Control Skills of Children.

This section presents the findings related to the research question “*Do the academic self-control skills of the children whose parents participated in the training improve significantly after the training?*”. The findings from the analysis conducted on the total scores, academic perseverance and academic attention sub-scales obtained from the ASCS applied to children before and after the training, as well as the scores related to the total and sub-scales of the ASCS for the children in the experimental group during a three-month follow-up are given. Descriptive statistics for the data collected regarding academic self-control and its sub-scales are provided in Table 2.

**Table 2.** ASCS descriptive statistics.

	Experimental group ( $n = 11$ )						Control group ( $n = 11$ )			
	Pre-test		Post-test		Follow up		Pre-test		Post-test	
	$\bar{X}$	$SD$	$\bar{X}$	$SD$	$\bar{X}$	$SD$	$\bar{X}$	$SD$	$\bar{X}$	$SD$
Academic perseverance	29.45	7.11	35.09	4.15	33.54	5.08	28.81	4.97	28.63	4.78
Academic attention	13.90	5.30	17.45	3.67	17.90	2.38	15.81	2.35	16.09	2.11
Academic self-control total	43.36	10.35	52.54	6.80	51.45	7.03	44.72	6.54	44.90	6.17

Table 2 shows that the average pre-test total score of the students in the ASCS for the experimental and control groups was 43.36 ( $SD = 10.35$ ) and 44.72 ( $SD = 6.17$ ), respectively. Post-test scores were 52.54 ( $SD = 6.80$ ) for the experimental group and 44.90 ( $SD = 6.17$ ) for the control group. The average ASCS score for the experimental group after a three-month follow-up was 51.45 ( $SD = 7.03$ ). Regarding the sub-dimension scores, the average scores for academic perseverance in the experimental group were 29.45 ( $SD = 7.11$ ) for the pre-test, 35.09 ( $SD = 4.15$ ) for the post-test, and 33.54 ( $SD = 5.08$ ) for the follow-up. In the control group, the average score for academic perseverance scores were 28.81 ( $SD = 4.97$ ) for the pre-test and 28.63 ( $SD = 4.78$ ) for the post-test. The average academic attention scores for the experimental group was 13.90 ( $SD = 5.30$ ) in the pre-test, 17.45 ( $SD = 3.67$ ) in the post-test and 17.90 ( $SD = 2.38$ ) in the follow-up. In the control group, the result of the pre-test was 15.81 ( $SD = 2.35$ ) and the result of the post-test was 16.09 ( $SD = 2.11$ ). The results show that the average score of the experimental group increased from the pre-test to the post-test and the score at the three-

month follow-up remained higher compared to the pre-test. In the control group, there was a limited increase in the average scores for academic attention and total scores from the pre-test to post-test, except for academic perseverance. To determine whether these increases were statistically significant, the differences in ASCS pre-test total and sub-dimension scores between the groups were analyzed using the Mann-Whitney U test. The results are presented in Table 3.

**Table 3.** Mann-Whitney U Test results for pre-test and post-test ASCS scores by groups.

Process	Dimension	Group	<i>n</i>	Mean Rank	Total Rank	<i>U</i>	<i>p</i>	<i>r</i>
Pre-test	Academic perseverance	Experimental	11	15.27	168.00	55.00	.748	-
		Control	11	7.73	85.00			
	Academic attention	Experimental	11	11.00	121.00	51.00	.562	-
		Control	11	12.00	132.00			
	Academic self-control total	Experimental	11	11.45	126.00	60.00	.100	-
		Control	11	11.55	127.00			
Post-test	Academic perseverance	Experimental	11	15.27	168.00	19.00	.005	.58
		Control	11	7.73	85.00			
	Academic attention	Experimental	11	13.82	152.00	35.00	.101	-
		Control	11	9.18	101.00			
	Academic self-control total	Experimental	11	14.64	161.00	26.00	.023	.48
		Control	11	8.36	92.00			

When examining the findings of the Mann-Whitney U-test presented in Table 3, it is observed that there is no significant difference between the groups on the pre-test measures for either the total scores of academic self-control ( $U = 60$ ;  $p > .05$ ) or the sub-dimensions of academic perseverance ( $U = 55$ ;  $p > .05$ ) and academic attention ( $U = 51$ ;  $p > .05$ ). This indicates that the level of academic self-control was similar among the students before the program. However, the findings from the analysis of whether the post-test total and sub-dimension scores of the ASCS differentiated between the groups after the program indicate a difference between the groups. A significant difference was found between the groups in terms of the total score of the ASCS ( $U = 26$ ;  $p < .05$ ;  $r = .58$ ) and the academic perseverance dimension ( $U = 19$ ;  $p < .05$ ;  $r = .48$ ), whereas no significant difference was observed in the academic attention dimension ( $U = 35$ ;  $p > .05$ ). Considering the fact that the reliability of the academic attention sub-dimension was lower compared to other dimensions during the development of ASCS, the weakness of the items in measuring this variable may have led to this result. It is therefore assumed that there was no significant difference between the groups in this dimension. The results of the Wilcoxon Signed Ranks test, which shows the differentiation of the children's academic self-control scores before and after the training, are shown in Table 4.

Table 4 shows that there was a significant difference in the experimental group between the total academic self-control score ( $Z = -2.81$ ;  $p < .05$ ;  $r = .85$ ), the academic perseverance dimension ( $Z = -2.81$ ;  $p < .05$ ;  $r = .85$ ), and the academic attention dimension ( $Z = -2.81$ ;  $p < .05$ ;  $r = .85$ ) before and after training. The effect size for the difference in each of the three score types was calculated as .85 ( $r = z / \sqrt{N}$ ), and this effect was interpreted as large according to Cohen's (1988) suggestion. When looking at the median values, there was an increase in the experimental group in both the ASCS total score (median [ $Mdn$ ] = 47 pre-test,  $Mdn = 55$  post-test), the academic perseverance sub-dimension score ( $Mdn = 16$  pre-test,  $Mdn = 19$  post-test) and the academic attention sub-dimension score ( $Mdn = 32$  pre-test,  $Mdn = 36$  post-test). The difference observed in the experimental group favors the results of the post-test. It can be



concluded that the academic self-control skills of the children whose mothers participated in the training improved significantly after the training.

**Table 4.** Wilcoxon Signed Ranks test results for pre-test and post-test the ASCS scores by groups.

Subscale	Group	Ranks	<i>n</i>	Mean Rank	Total Rank	<i>Z</i>	<i>p</i>	<i>r</i>
Academic perseverance	Experimental	Negative rank	1	1.50	1.50	-2.81	.005	.85
		Positive rank	10	6.45	64.50			
		Equal	0					
	Control	Negative rank	6	5.58	33.50	-0.05	.964	-
		Positive rank	5	6.50	32.50			
		Equal	0					
Academic attention	Experimental	Negative rank	1	0.00	0.00	-2.81	.005	.85
		Positive rank	10	5.50	55.00			
		Equal	0					
	Control	Negative rank	2	3.00	6.00	-0.10	.317	-
		Positive rank	4	3.75	15.00			
		Equal	5					
Academic self-control total	Experimental	Negative rank	0	0.00	0.00	-2.81	.005	.85
		Positive rank	10	5.50	55.00			
		Equal	1					
	Control	Negative rank	4	3.75	15.00	-0.42	.672	-
		Positive rank	4	5.25	21.00			
		Equal	3					

In the control group, there was no significant difference between the pre-test and post-test total ASCS scores ( $Mdn = 46$  pre-test,  $Mdn = 46$  post-test) and the sub-dimension scores (for academic perseverance,  $Mdn = 16$  pre-test,  $Mdn = 16$  post-test; for academic attention,  $Mdn = 29$  pre-test,  $Mdn = 28$  post-test) ( $p > .05$ ). When examining the mean rank values of the groups for the overall assessment of academic self-control, academic perseverance and academic attention, it can be seen that the differences in the experimental group are greater than in the control group. The academic self-control scores of the children whose mothers participated in the program increased compared to those whose mothers did not.

The findings of the Friedman test performed to compare the total and subscale scores of three different measures of the ASCS of the experimental group are shown in Table 5. The analysis results show a statistically significant difference in the pre-test, post-test, and three-month follow-up total scores of the children in the experimental group (Fr.  $\chi^2_{(2, 11)} = 15.95, p = .000$ ). After Bonferroni correction in the post hoc analysis, this difference between the pre-test and the post-test was found to be significant ( $p = .000$ ). The effect size of this difference was calculated as .73 and interpreted as a large effect size. Accordingly, the post-test ASCS scores of the children in the experimental group are significantly higher than their pre-test scores. No significant difference was found between the pre-test and follow-up ( $p = .076$ ) and between the post-test and follow-up measurements ( $p = .329$ ). These findings indicate that the children's academic self-control levels increased from the pre-test to the post-test following the training, and this increase continued during the follow-up. When examining the findings related to the academic attention variable, the difference between the three different measurements of the children's academic attention scores in the experimental group is significant, and the effect size was calculated as .58 (Fr.  $\chi^2_{(2, 11)} = 13.85, p = .001$ ).

In the pairwise comparisons with Bonferroni correction, this difference was found to be significant between the pre-test and the post-test ( $p = .003$ ) and between the pre-test and the

follow-up ( $p = .017$ ). The lack of a significant difference between the post-test and the three-month follow-up scores ( $p = 1.00$ ) indicates that the increase continued in the same manner, demonstrating the ongoing effectiveness of the program implemented. Looking at the median scores, it can be seen that the pre-test academic attention scores ( $Mdn = 16$ ) of the children in the experimental group are lower than the post-test ( $Mdn = 19$ ) and three-month follow-up scores ( $Mdn = 19$ ). These findings indicate that the academic attention of the children whose mothers participated in the training increased significantly after the training. The academic perseverance scores of the children in the experimental group showed a statistically significant difference in the three different measurements (Fr.  $\chi^2_{(2, 11)} = 12.66, p = .002$ ). In the pairwise comparisons, this difference was found to be between the pre-test and post-test ( $p = .003$ ), while no significant difference was found between the pre-test and follow-up ( $p = .099$ ) and between the post-test and follow-up ( $p = .723$ ). The calculation of the effect size ( $W = \chi^2/n (K-1)$ ) indicates a large effect size (.63).

**Table 5.** Friedman Test findings for the experimental group ASCS pre-test, post-test, and follow-up.

							Difference b/w groups ( <i>p</i> )		
Process	<i>n</i>	<i>Mdn</i>	<i>df</i>	$\chi^2$	<i>p</i>	<i>W</i>	Pre-Post	Pre-Follow	Post-Follow
Academic self-control total									
Pre-test	11	47	2	15.95	.000	.73	.000	.076	.329
Post-test	11	55							
Follow-up	11	54							
Academic attention									
Pre-test	11	16	2	13.85	.001	.63	.033	.017	1.000
Post-test	11	19							
Follow-up	11	19							
Academic perseverance									
Pre-test	11	32	2	12.66	.002	.58	.003	.099	.723
Post-test	11	36							
Follow-up	11	36							

Like a muscle, self-control can be developed through exercise (Baumeister *et al.*, 2007). Academic self-control is also a trainable skill and can be improved with practice. Therefore, there are numerous trainings, exercises, and programs aim to develop self-control. These can be delivered directly to individuals or indirectly through parents and teachers (Duckworth *et al.*, 2014). Reviews of the literature examining the relationship between self-control and parenting suggest that most studies are predominantly conducted in early childhood and adolescence, are problem-focused (emotional and behavioral problems, depression, addictions, etc.), and include specific groups (e.g., those with attention deficit hyperactivity disorder, eating disorders, anxiety, etc.). For example in a longitudinal study examining the relationship between maternal self-regulation and child self-regulation (Bolger *et al.*, 2022), data were collected on mothers' self-regulation when the children were 6 months old, maternal attachment when the children were 7 years old, parenting practices at age 12, and adolescents' low self-regulation at age 15. Using structural equation modeling, the study found that lower maternal self-regulation was positively associated with lower adolescent self-regulation through its effects on maternal attachment and, subsequently, parenting behaviors. Common findings suggest that inappropriate parenting practices are associated with low parental self-control. Similarly, low parental self-control is associated with low child self-control. Low self-control

is seen as both a cause and a consequence of negative parenting (Bolger *et al.*, 2022; Coyne & Wright, 2014; Cullen *et al.*, 2008; Ng-Knight *et al.*, 2016).

### 3.2. Findings Related to the Parenting Competence Of Mothers

Findings related to the research question “Do the parenting competence of the parents who participated in the training improve significantly after the training?” are given in this section. Findings and comments regarding the pre-test and post-test total scores from the PSCS of mothers who participated in the parent training and mothers who did not, as well as the total scores from the three-month follow-up test of mothers from the experimental group, are presented. The descriptive statistics for the pre-test, post-test and three-month follow-up measurements of the PSCS scores of the groups are presented in Table 6.

**Table 6.** Descriptive statistics of parenting competence.

Process	Experimental group			Control group		
	<i>n</i>	$\bar{X}$	<i>SD</i>	<i>n</i>	$\bar{X}$	<i>SD</i>
Pre-test	11	65.09	12.05	11	61.36	7.04
Post-test	11	73.00	10.62	11	61.63	7.03
Follow up	11	72.09	11.94	-	-	-

As shown in Table 6, the pre-test PSCS mean scores for mothers in the experimental and control groups are 65.09 ( $SD = 12.05$ ) and 61.36 ( $SD = 7.04$ ), respectively, while the post-test mean scores are 73.00 ( $SD = 10.62$ ) and 61.63 ( $SD = 7.03$ ), respectively. The mean follow-up score for the experimental group was 72.09 ( $SD = 11.94$ ). The post-test mean scores of mothers in the experimental and control groups appear to have increased compared to their pre-test mean scores. However, this increase appears to be higher in the experimental group. Although the mean follow-up scores of the experimental group are slightly lower than the post-test mean scores, they are still higher than the pre-test mean scores. The Mann-Whitney U test was performed to determine whether these differences were statistically significant and its findings are shown in Table 7.

**Table 7.** Mann-Whitney U Test Results for pre-test and post-test PSCS scores by groups.

Process	Group	<i>n</i>	Mean Rank	Total Rank	<i>U</i>	<i>p</i>	<i>r</i>
Pre-test	Experimental	11	12.27	135.00	52.00	.606	-
	Control	11	10.73	118.00			
Post-test	Experimental	11	14.95	164.50	22.50	.010	.53
	Control	11	8.05	88.50			

According to the results of the Mann-Whitney U-test on the mothers' pre-test and post-test PSCS total scores, shown in Table 7, there is no significant difference between the groups in the pre-test PSCS total scores ( $U = 52.00$ ;  $p > .05$ ). On the other hand, there is a significant difference between the groups in post-test total scores ( $U = 22.50$ ;  $p < .05$ ). According to the effect size calculation (Cohen, 1988), this difference was calculated as .53.

The Wilcoxon Signed Ranks Test was conducted to determine the difference between the pre-test and post-test PSCS total scores in the experimental and control groups. The results of this analysis are shown in Table 8. As seen in Table 8, there is a significant difference between the pre-test ( $Mdn = 60$ ) and post-test ( $Mdn = 73$ ) total PSCS scores of mothers who received training ( $Z = -2.60$ ,  $p < .05$ ,  $r = .78$ ). In the group that did not receive training, however, there is no significant difference between the pre-test ( $Mdn = 63$ ) and post-test ( $Mdn = 62$ ) total PSCS scores ( $Z = -0.72$ ,  $p > .05$ ). These findings suggest that the PSCS post-test scores of mothers who took the training increased significantly. According to Cohen (1988), an  $r$  value between .3 and .5 indicates a medium effect, while a value greater than .5 indicates a strong effect. The



difference created by the implemented program on the parenting competence of the mothers can be considered a strong effect. In contrast, no significant change was observed between the pre-test and post-test scores of the mothers in the control group.

**Table 8.** Wilcoxon Signed Ranks test results for pre-test and post-test the PSCS scores by groups.

Group	Ranks	<i>n</i>	Mean Rank	Total Rank	<i>Z</i>	<i>p</i>	<i>r</i>
Experimental	Negative rank	1	3.50	3.50	-2.60	.009	.78
	Positive rank	10	6.25	62.50			
	Equal	0					
Control	Negative rank	3	4.33	13.00	-0.72	.470	-
	Positive rank	5	4.60	23.00			
	Equal	3					

The findings of the Friedman Analysis conducted to test the differences between the three different PSCS measurements of the mothers in the experimental group are presented in Table 9. As seen in Table 9, the results of the Friedman Test conducted to observe the differences between the three different measurements of the experimental group show a statistically significant difference in the total PSCS scores of the mothers across the three time points (Fr.  $\chi^2_{(2, 11)} = 7.42, p = .024$ ). Bonferroni correction was applied for post-hoc analyses. The findings indicate that the difference between the measurements is significant between the pre-test and post-test. The effect size of the difference was calculated as .34, which, according to Cohen (1988), is interpreted as a medium effect. Thus, the PSCS post-test scores of mothers in the experimental group are significantly higher than their pre-test scores ( $p = .032$ ). However, the differences between the pre-test and follow-up scores ( $p = .165$ ) and the post-test and follow-up scores ( $p = 1.00$ ) were not found to be significant ( $p > .05$ ). When examining the median values of the measurements, it can be seen that there is an increase in scores from the pre-test ( $Mdn = 60$ ) to the post-test ( $Mdn = 73$ ), and the three-month follow-up scores ( $Mdn = 72$ ) are still higher than the pre-test scores. This suggests that the parenting competence levels of the mothers increased after the training.

**Table 9.** Friedman Test findings for the experimental group PSCS pre-test, post-test, and follow-up.

Process	<i>n</i>	<i>Mdn</i>	<i>df</i>	$\chi^2$	<i>p</i>	<i>W</i>	Difference b/w groups ( <i>p</i> )		
							Pre-Post	Pre-Follow	Post-Follow
Pre-test	11	60							
Post-test	11	73	2	7.42	.024	.34	.032	.165	1.000
Follow-up	11	72							

Feeling competent and confident in one's parenting roles is defined as parenting competence (Coleman & Karraker, 2000). Parenting competence is considered a key determinant of parenting behaviors (Hill & Bush, 2001; Jones & Prinz, 2005). Therefore, enhancing parenting competence is a common objective of family education programs (Haslam *et al.*, 2016). Numerous studies have shown that such programs are effective in strengthening parenting competence (Kaminski *et al.*, 2008; Sanders *et al.*, 2014). Within these programs, the support and skills parents acquire help reinforce their belief in and confidence about their parenting, ultimately contributing to an increase in positive parenting behaviors.

#### 4. DISCUSSION and CONCLUSION

The aim of this study was to contribute to the development of children's academic self-control skills through positive parenting. To achieve this goal, a program called FSASC was developed to provide parents with suggestions and practical strategies to support their children

academically. The program included practices such as emotion validation to enhance communication, establishing family rules, setting boundaries, and creating routines primarily regarding internet use and eating habits. Real-life scenarios were used throughout the six sessions to make the program practical and relatable for parents. During the program development process, the challenges and support needs most commonly experienced by families regarding their children's school life were identified and prioritized.

In this experimental study, mothers participated in the online training weekly for eight weeks. Their parental competence, as well as their children's academic self-control, were compared to those of the control group. The findings showed that both the mothers who participated in the program and their children benefited from the FSASC. Children of mothers who received the training showed a significant increase in academic self-control and academic perseverance scores compared to those whose mothers did not receive the training. This improvement was maintained at the three-month follow-up. However, no significant difference was found between the groups in the academic attention subscale. This result may be due to the relatively low internal consistency of this subscale ( $\alpha = .636$ ), indicating limited item reliability for capturing the construct accurately.

The program also led to an increase in the mothers's parenting skills, and this improvement was still observed at three-month follow-up. It can be said that the mothers and children who participated in the training benefited from the FSASC compared to the control group. The academic self-control levels of the children and the parenting competencies of the mothers increased significantly. After the program, the mothers expressed that they had benefited from the training and were satisfied with their participation.

During the pandemic, under remote education conditions, distractions became stronger than ever, and external control decreased. As a result, higher levels of self-control were required to fulfill academic responsibilities. After this relatively less structured period of remote learning, students began to report difficulties in managing academic demands. These students not only fell behind in schoolwork but also in extracurricular activities and social interaction with peers, making it more difficult to demonstrate academic self-control. Meanwhile, both parents and teachers expressed concern about students who avoided studying, missed lessons, and neglected academic responsibilities. Therefore it seems both important and necessary to teach parents a set of practices that they can implement to help their children assume and fulfill their academic responsibilities independently. When children deal with their school-related works on their own, it not only supports academic success but also reduces family-child conflict.

Some practical suggestions can be made for how these findings might be applied in educational settings. According to a comparative study conducted by the OECD (2021), 10-year-old students demonstrated significantly higher levels of social and emotional skills compared to 15-year-olds. In that study, self-control was assessed under the broader domain of task performance, alongside responsibility and perseverance. These findings indicate that the pre-adolescent period is a particularly critical time for developing and reinforcing self-control skills. Therefore, school-based interventions should prioritize this developmental window to cultivate students' capacity for behavioral regulation and academic persistence. Embedding self-control strategies into daily classroom routines, such as goal setting, planning, and self-reflection, can help students internalize effective academic habits. Teachers, just like parents, can play an essential role by providing immediate and constructive feedback and encouraging students to monitor their attention, effort, and emotional responses. All these habits and skills positively affect academic success. Therefore, programs aimed at supporting academic self-control may be developed and implemented by involving both teachers and schools.

Feeling parental supervision and support, and growing up in a well-structured, safe, and orderly environment, contributes to the early development of self-control. Behavioral self-control begins to develop in early childhood and is largely influenced by parenting strategies such as

praise, encouragement, warmth, and guidance (Calkins *et al.*, 2002). It is also known that the use of negative parenting strategies predicts lower levels of self-control in children (Beaver *et al.*, 2010; Cullen *et al.*, 2008). When discussing parenting, it is more meaningful to consider both mothers and fathers together, if possible. However, in this study, only mothers participated, and fathers were not included. Therefore, this can be considered a limitation of the study. Future studies could involve mixed groups that include both mothers and fathers, or focus specifically on fathers to examine the effectiveness of the training from a broader parental perspective.

In this study, the parent training program was implemented remotely and individually due to the Covid-19 pandemic conditions. The online format had both positive and negative effects on its overall effectiveness. On the positive side, the absence of time and location constraints was a factor in increasing maternal participation. For example, scheduling sessions after 9 p.m. for working mothers, or rescheduling for those who missed a session due to unexpected commitments, proved to be effective. This flexibility -allowing missed sessions to be easily made up- was considered helpful, based on verbal feedback from both facilitators and participants. On the other hand, the lack of dynamism typical of face-to-face training, limited physical interaction in activities such as role-playing, and occasional technical issues (e.g., sound or video problems due to internet connectivity) were seen as disadvantages of online implementation, and as a limitation of the study. It is recommended that the effectiveness of the program also be evaluated in a face-to-face and group setting.

Additionally, this study was conducted in a residential area with families from low- to middle-income backgrounds, all enrolled in public schools. The impact of the program may also be examined among different socioeconomic groups, including private school populations. Future academic self-control studies may involve not only parents but also their children in the intervention process. Similar programs can also be tested in school settings by including classroom guidance counselors in middle schools and classroom teachers in primary schools.

In all areas of child development, parents play a central supportive role. They influence their children positively or negatively not only through genetic inheritance but also through their parenting practices and the environment they create. For this reason, any effort directed toward supporting parents is also an investment in child development. Like other studies focusing on parents, this research can be considered a contribution to children's overall well-being.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Ankara University Social Sciences Ethics Committee, 13.03.2020-3/58.

### Contribution of Authors

**Hanife Büyük:** Literature review, Resources, Methodology, Data collection, Data analysis, Writing, and Editing. **Selahiddin Öğülmüş:** Supervision and Validation.

### Orcid

Hanife Büyük  <https://orcid.org/0000-0002-9437-0656>

Selahiddin Öğülmüş  <https://orcid.org/0000-0002-8737-5141>

### REFERENCES

- Baumeister, R.F., & Vohs, K.D. (2007). Self regulation, ego depletion and motivation. *Social and Personality Psychology Compass*, 1(1), 115-128.
- Baumeister, R., Vohs, K., & Tice, D. (2007). The strength model of self-control. *Current Directions in Psychological Science*. 16, 351-355. <https://doi.org/10.1111/j.1467-8721.2007.00534.x>

- Beaver, K.M., Ferguson, C.J., & Lynn-Whaley, J. (2010). The association between parenting and levels of self-control: A genetically informative analysis. *Criminal Justice and Behavior*, 37(10), 1045-1065. <https://doi.org/10.1177/0093854810374919>
- Bolger, M.A., Meldrum, R.C., & Liu, L. (2022). Maternal low self-control, maternal attachment toward children, parenting practices, and adolescent low self-control: A prospective 15-year study. *Journal of Developmental and Life-Course Criminology*, 8, 206-231. <https://doi.org/10.1007/s40865-022-00198-8>
- Bozgün, K., & Başgöl, M. (2018). Akademik Azim Ölçeği'nin Türkçe'ye uyarlanması: Geçerlik ve güvenirlik çalışması [Adaptation of the Academic Grit Scale into Turkish: A study on validity and reliability]. *Journal of Academic Social Research*, 6(85), 435-445. <https://doi.org/10.16992/ASOS.14521>
- Büyük, H., Öğülmüş, S., & Kapçı, E.G. (2020). Academic self-control scale for secondary school students: Validity and reliability study. *Turkish Journal of Education*, 9(4), 290-306. <https://doi.org/10.19128/turje.778117>
- Büyüköztürk, Ş. (2009). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum*. [Data analysis handbook for social sciences: Statistics, research design, SPSS applications and interpretation] (10th ed.). Pegem A Publication.
- Calkins, S.D., Dedmon, S.E., Gill, K.L., Lomax, L.E., & Johnson, L.M. (2002). Frustration in infancy: Implications for emotion regulation, physiological processes, and temperament. *Infancy*, 3(2), 175-197. [https://doi.org/10.1207/S15327078IN0302\\_4](https://doi.org/10.1207/S15327078IN0302_4)
- Chen, J.J.L. (2005). Relation of academic support from parents, teachers, and peers to Hong Kong adolescents' academic achievement: The mediating role of academic engagement. *Genetic, Social, and General Psychology Monographs*, 131(2), 77-127. <https://doi.org/10.3200/MONO.131.2.77-127>
- Cho, Y.I., Kim, J.S., & Kim, J.O. (2018). Factors influencing adolescents' self-control according to family structure. *Journal of Child and Family Studies*, 27(11), 3520-3530. <https://doi.org/10.1007/s10826-018-1175-4>
- Christensen, L.B., Johnson, R.B., & Turner, L.A. (2015). *Araştırma yöntemleri: Desen ve analiz* (A. Aypay, Trans. Ed.) [Research methods: Design and analysis]. Anı Publishing.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.
- Coleman, P.K., & Karraker, K.H. (2000). Parenting self-efficacy among mothers of school-age children: Conceptualization, measurement, and correlates. *Family Relations: An Interdisciplinary Journal of Applied Family Studies*, 49(1), 13-24. <https://doi.org/10.1111/j.1741-3729.2000.00013.x>
- Coyne, M.A., & Wright, J.P. (2014). The stability of self-control across childhood. *Personality and Individual Differences*, 69, 144-149. <https://doi.org/10.1016/j.paid.2014.05.026>
- Cullen, F.T., Unnever, J.D., Wright, J.P., & Beaver, K.M. (2008). Parenting and self-control. In Goode, E. (Ed.), *Out of control: Assessing the general theory of crime* (pp. 61-74). Stanford University Press.
- Dishion, T.J., & McMahon, R.J. (1998). Parental monitoring and the prevention of child and adolescent problem behavior: A conceptual and empirical formulation. *Clinical Child and Family Psychology Review*, 1(1), 61-75. <https://doi.org/10.1023/A:1021800432380>
- Duckworth A.L., Gendler T.S., & Gross J.J. (2014). Self-control in school-age children. *Educational Psychologist*, 49(3), 199-217. <https://doi.org/10.1080/00461520.2014.926225>
- Duckworth, A.L., & Seligman, M.P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16(12), 939-944. <https://doi.org/10.1111/j.1467-9280.2005.01641.x>



- Duckworth, A.L., Taxer, J.L., Eskreis-Winkler, L., Galla, B.M., & Gross, J.J. (2019). Self-control and academic achievement. *Annual Review Psychology*, 70, 373-399. <https://doi.org/10.1146/annurev-psych-010418-103230>
- Feinstein, L., Duckworth, K., & Sabates, R. (2008). *Education and the family: Passing success across the generations* (1st ed.). Routledge.
- Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.
- Finkenauer C., Engels R., & Baumeister R. (2010). Parenting behavior and adolescent behavioral emotional problems: The role of self-control. *International Journal of Behavioral Development*, 29(1), 58-69. <https://doi.org/10.1080/01650250444000333>
- Galla, B.M., & Duckworth, A.L. (2015). More than resisting temptation: Beneficial habits mediate the relationship between self-control and positive life outcomes. *Journal of Personality and Social Psychology*, 109(3), 508-525. <https://doi.org/10.1037/pspp0000026>
- Galla, B.M., Amemiya, J., & Wang, Ming-T. (2018). Using expectancy-value theory to understand academic self-control. *Learning and Instruction*, 58, 22-33. <https://doi.org/10.1016/j.learninstruc.2018.04.004>
- Gottfredson, M.R., & Hirschi, T. (1990). *A general theory of crime*. Stanford University Press.
- Gottfried, A.E., Fleming, J.S., & Gottfried, A.W. (1994). Role of parental motivational practices in children's academic intrinsic motivation and achievement. *Journal of Educational Psychology*, 86, 104-113. <https://doi.org/10.1037//0022-0663.86.1.104>
- Haslam, D., Mejia, A., Sanders, M., & de Vries, P. (2016). Parenting programs. In J. M. Rey (Ed.), *IACAPAP e-textbook of child and adolescent mental health* (pp. 1–29). International Association for Child and Adolescent Psychiatry and Allied Professions.
- Hill, N., & Bush, K. (2001). Relationships between parenting environment and children's mental health among African American and European American mothers and children. *Journal of Marriage and The Family*, 63, 954-966. <https://doi.org/10.1111/j.1741-3737.2001.00954.x>
- Jones, T.L., & Prinz, R.J. (2005). Potential roles of parental self-efficacy in parent and child adjustment: A review. *Clinical Psychology Review*, 25, 341-363. <https://doi.org/10.1016/j.cpr.2004.12.004>
- Kabat-Zinn, J. (2003). Mindfulness-Based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice*, 10, 144-156.
- Kağıtçıbaşı, Ç. (2005). Autonomy and relatedness in cultural context: Implications for self and family. *Journal of Cross-Cultural Psychology*, 36(4), 403-422. <https://doi.org/10.1177/0022022105275959>
- Kaminski, J.W., Valle, L.A., Filene, J.H., & Boyle, C.L. (2008). A meta-analytic review of components associated with parent training program effectiveness. *Journal of Abnormal Child Psychology*, 36(4), 567–589. <https://doi.org/10.1007/s10802-007-9201-9>
- Kennett, D.J. (1994). Academic self-management counselling: Preliminary evidence for the importance of learned resourcefulness on program success, *Studies on Higher Education*, 19(3), 295-307.
- Kennett, D.J., & Maki, K. (2014, April 28). *Academic resourcefulness and transfer student success: Direct entry, college transfer, and university transfer student comparisons* [Conference presentation]. ONCAT Student Pathways Conference. [https://oncat.ca/sites/default/files/inline-images/1b\\_2014.pdf](https://oncat.ca/sites/default/files/inline-images/1b_2014.pdf)
- Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H., ... Sears, M.R. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7), 2693-2698. <https://doi.org/10.73/pnas.1010076108>

- Muraven, M., & Baumeister, R.F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, 126, 247–259.
- Ng-Knight, T., Shelton, K., Riglin, L., Mcmanus, I., Frederickson, N., & Rice, F. (2016). A longitudinal study of self-control at the transition to secondary school: Considering the role of pubertal status and parenting. *Journal of Adolescence*, 50, 44-55. <https://doi.org/10.1016/j.adolescence.2016.04.006>
- O’Connell, M.E., Boat, T., & Warner, K.E. (2009). *Preventing mental, emotional and behavioral disorders among young people: Progress and possibilities*. National Academies Press.
- OECD. (2021). *Social and emotional skills: Turkish national preliminary report* [No. 19]. Ministry of National Education, Republic of Türkiye. [https://www.meb.gov.tr/meb\\_iys\\_dosyalar/2021\\_09/07170836\\_No19\\_-\\_OECD\\_Sosyal\\_ve\\_Duygusal\\_Beceriler\\_Arastirmasi.pdf](https://www.meb.gov.tr/meb_iys_dosyalar/2021_09/07170836_No19_-_OECD_Sosyal_ve_Duygusal_Beceriler_Arastirmasi.pdf)
- Piquero, A.R., Jennings, W.G., Diamond, B., Farrington, D.P., Tremblay, R.E., Welsh, B.C., & Gonzalez, J.M.R. (2016). A meta-analysis update on the effects of early family/ parent training programs on antisocial behavior and delinquency. *Journal of Experimental Criminology*, 12(2), 229-248. <https://doi.org/10.1007/s11292-016-9256-0>
- Rebecca, D.M. (2015). *To be kind or not to be kind: The role of self-compassion in the academic self-control model* [Unpublished master’s thesis]. Trent University.
- Rebecca D.M., & Kennett, D.J. (2018). To be kind or not to be kind: The moderating role of self-compassion in the relationship between general resourcefulness and academic self-regulation. *The Journal of Social Psychology*, 158(5), 626-638. <https://doi.org/10.1080/00224545.2017.1407286>
- Robson, D.A., Allen, M.S. & Howard, S.J. (2020). Self-regulation in childhood as a predictor of future outcomes: A meta-analytic review. *Psychological Bulletin*, 146(4), 324-354. <https://doi.org/10.1037/bul0000227>
- Rosenbaum, M. (1990). The role of learned resourcefulness in the self-control of health behavior. In M. Rosenbaum (Ed.). *Learned resourcefulness: On coping skills, selfcontrol, and adaptive behavior* (pp. 3-30). Springer Publishing Co.
- Rosenbaum, M. (2000). The self-regulation of experience: Openness and construction. In P. Dewe, A.M. Leiter, & T. Cox (Eds.), *Coping and health in organizations* (pp. 51–67). Taylor and Francis.
- Sanders, M.R., Kirby, J.N., Tellegen, C.L., & Day, J.J. (2014). The Triple P-Positive Parenting Program: A systematic review and meta-analysis of a multi-level system of parenting support. *Clinical Psychology Review*, 34(4), 337-357. <https://doi.org/10.1016/j.cpr.2014.04.003>
- Smithers, L.G., Sawyer, A.C.P., Chittleborough, C.R., Davies, N.M., Davey Smith, G., & Lynch, J.W. (2018). A systematic review and meta-analysis of effects of early life non-cognitive skills on academic, psychosocial, cognitive and health outcomes. *Nature Human Behaviour*, 2, 867–880. <https://doi.org/10.1038/s41562-018-0461-x>
- Sroufe, A.L. (2005). Attachment and development: a prospective, longitudinal study from birth to adulthood. *Attachment and Human Development*, 7(4), 349-367. <https://doi.org/10.1080/14616730500365928>
- Steinberg, L. (2018). *Fırsatlar dönemi olarak ergenlik [Age of Opportunity: Lessons from the New Science of Adolescence]* (E. Boynueğri, Trans.) . İmge Publishing. (Original work published 2014)
- Strayhorn, J.M. (2002). Self-control: Toward systematic training programs. *Journal of the American Academy of Child and Adolescent Psychiatry*, 41(1), 17-27. <https://doi.org/10.1097/00004583-200201000-00007>

- Tangney, J.P., Baumeister, R.F., & Boone, A.L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality and Social Psychology*, 72(2), 271-324. <https://doi.org/10.1111/j.0022-3506.2004.00263.x>
- Vazsonyi, A.T., & Huang, L. (2010). Where self-control comes from: On the development of self-control and its relationship to deviance over time. *Developmental Psychology*, 46(1), 245-257. <https://doi.org/10.1037/a0016538>
- Vazsonyi, A.T., Javakhishvili, M., & Blatny, M. (2021). Does self-control outdo IQ in predicting academic performance? *Journal of Youth and Adolescence*, 51(3), 499-508. <https://doi.org/10.1007/s10964-021-01539-4>
- Vygotsky, L.S. (1984). Interaction between learning and development. In L.S. Vygotsky, *Mind in society: The development of higher psychological processes* (pp. 79–91). Harvard University Press. (Original work published 1978).
- Williamson, L.Z. (2021). *Becoming Odysseus: Using situational and cognitive self-control to attain academic goals in daily life* [Unpublished doctoral dissertation]. University of Wyoming.
- Zahn-Waxler, C., Klimes-Dougan, B., & Slattery, M.J. (2000). Internalizing problems of childhood and adolescence: Prospects, pitfalls, and progress in understanding the development of anxiety and depression. *Development and Psychopathology*, 12, 443-466. <https://doi.org/10.1017/S0954579400003102>

## Mathematical literacy, thinking, and teacher self-efficacy: A latent profile analysis of preservice teachers' competencies

Muhammed Celal Uras<sup>1</sup>, Mehmet Şata<sup>2\*</sup>, Yasin Soylu<sup>3</sup>

<sup>1</sup>Ağrı İbrahim Çeçen University, Faculty of Education, Department of Mathematics Education, Ağrı, Türkiye

<sup>2</sup>Van Yüzüncü Yıl University, Faculty of Education, Department of Educational Sciences, Van, Türkiye

<sup>3</sup>Atatürk University, Faculty of Kazım Karabekir Education, Department of Mathematics Education, Erzurum, Türkiye

### ARTICLE HISTORY

Received: Sep. 23, 2024

Accepted: Aug. 2, 2025

### Keywords:

Preservice teachers,  
Latent profile analysis,  
Professional  
competencies.

**Abstract:** Professional competence is an important factor in determining the quality of teachers' education. A limited number of studies examine preservice teachers' professional competencies from a person-centered approach. The current study aims to identify latent profiles of preservice teachers' professional competencies. Data were collected from 545 preservice elementary mathematics teachers. The results showed significant relationships between self-efficacy in mathematical literacy, mathematical thinking, and teachers' self-efficacy in mathematical language. The findings revealed that preservice teachers were classified with three different profiles in professional competencies: low, medium, and high. The findings may help target preservice teachers in the risk group of professional competencies and develop programs that will enable them to develop these competencies.

## 1. INTRODUCTION

Mathematics is a universal basic need for individual development and social advancement. It also has an important role in shaping cognitive development, problem-solving ability, and logical reasoning skills (Cresswell & Speelman, 2020). In addition, it is necessary in many areas of daily life, from numerical skills to understanding scientific facts. Therefore, mathematics is essential for academic success and social advancement. Everyone who is involved in daily life needs effective mathematics education. Mathematics teachers play an important role in meeting this need. Teachers prepare individuals for the future professionally and socially by improving their mathematical skills (Schwarz & Kaiser, 2019). They also contribute to the training of knowledgeable and talented individuals. Therefore, mathematics education is considered important.

Teachers' professional competencies- one of the essential elements of mathematics education- play an important role in students' acquisition of mathematical skills and ensuring their learning (Uras *et al.*, 2024). In addition to effective communication skills, problem-solving ability, and

\*CONTACT: Mehmet ŞATA ✉ [mehmetsata@yyu.edu.tr](mailto:mehmetsata@yyu.edu.tr) 📍 Van Yüzüncü Yıl University, Faculty of Education, Department of Educational Sciences, Van, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>



classroom management, mathematics teachers must have mathematical literacy, mathematical thinking, and mathematical language self-efficacy (Uras *et al.*, 2025). These competencies improve the quality of mathematics education, enhance students' mathematical confidence, and foster more positive attitudes toward mathematics. Therefore, developing and supporting mathematics teachers' professional competencies is important.

### **1.1. The Importance of Mathematics Education and Mathematics Teachers**

Teaching mathematics encourages critical thinking, problem-solving, and logical reasoning. (Mukuka *et al.*, 2023; Peter, 2012). It also provides the opportunity to explain mathematical ideas and explore different problem-solving strategies through effective communication and collaboration (Nilimaa, 2023; Uras & Soylu, 2024). In addition, it has an important role in developing a positive attitude towards mathematics. Mathematics teachers are at the center of learning experiences and mathematical skill development. They guide the teaching of mathematical concepts and support meeting individual learning needs (Gardesten, 2023). Effective mathematics teachers create inclusive and supportive learning environments where students feel empowered to take risks and deal with mistakes and challenges. They collaborate with colleagues and follow current research and practices in mathematics education. In this way, they develop students' passion for mathematics and encourage them to succeed. Future mathematics teachers - preservice teachers - need a variety of knowledge and skills to ensure the quality of mathematics education and promote student success (Buchholtz *et al.*, 2023). These skills and knowledge, called professional competence, are necessary to guide instruction and support success. The Mathematical Knowledge for Teaching (MKT) Framework, developed by Ball *et al.* (2008), focuses on the specific mathematical knowledge teachers need to teach mathematics effectively. Mathematical content includes information about students' mathematical thinking, curriculum standards, and pedagogical strategies. MKT emphasizes the importance of teachers' in-depth understanding of mathematical concepts and their ability to translate them into effective teaching practices. In addition, Sociocultural Theory in the context of mathematics education - highlights the importance of language skills and communication in constructing mathematical knowledge (Steele, 2001).

Teachers who use mathematical language effectively can encourage mathematical thinking and develop a positive attitude toward mathematics. They create opportunities for students to express their mathematical ideas and discuss them with their peers. Additionally, the mathematical literacy framework has highlighted the importance of having mathematical literacy to communicate mathematical ideas and apply mathematical concepts in daily life. This also includes skills in interpreting and analyzing mathematical information (OECD, 2023). In order to realize the importance of mathematics in daily life, it is necessary to make connections with mathematical concepts. Therefore, preservice teachers must be well prepared to ensure the quality of mathematics education and to promote student achievement (Ekmekci & Serrano, 2022). Mathematics education equips students with numerical competence and develops interdisciplinary skills such as critical thinking, problem-solving, and logical reasoning. These skills have a wide range of effects, from academic performance to lifelong learning. Mathematics teachers play a key role in achieving learning and developing students' mathematical competencies. Therefore, training preservice teachers is critical to ensuring the quality of mathematics education and promoting student success. They need to be proficient in mathematical language to engage students effectively. They must have mathematical literacy and thinking skills to encourage mathematical thinking and develop a positive attitude toward mathematics (Yildiz *et al.*, 2020). The acquisition of these competencies by preservice teachers can enhance students' ability to develop self-confidence, mathematical competence, and enthusiasm for lifelong learning.

## 1.2. Mathematical Literacy Self-Efficacy, Mathematical Thinking and Teacher Self-Efficacy in Mathematical Language

Understanding the factors that influence students learning levels and achievements plays a decisive role in the education and training process, and some of it is related to teacher competencies (Jentsch & König, 2022). Among these factors, mathematics literacy self-efficacy (MLSE), mathematical thinking (MT), and teacher self-efficacy in mathematical language (TSEML) are of great interest because they are important in shaping students' mathematical experiences and outcomes. Therefore, research conducted in recent years has focused on examining the competencies of teachers and preservice teachers (Dervenis *et al.*, 2022). MLSE relates to an individual's beliefs about understanding, applying, and using mathematical concepts effectively in various contexts. It also includes confidence in their problem-solving skills, reasoning abilities, and communication proficiency (Ozgen, 2019). Preservice teachers' self-efficacy beliefs regarding mathematical literacy are high (Akkaya *et al.*, 2012). Moreover, studies examining preservice teachers' literacy self-efficacy beliefs have reported different results. Zehir and Zehir (2016) reported that preservice teachers' literacy self-efficacy beliefs differed significantly according to gender and grade level. On the other hand, Topbaş Tat (2018) reported that preservice teachers' literacy self-efficacy beliefs did not differ significantly according to gender and grade level. Those with high MLSE levels also have high interpretation, formulation, and problem-solving skills (Busnawir *et al.*, 2021). Teachers' MLSE levels are an important predictor of children's geometric shape recognition and numerical skills (Aktulun, 2018). Teachers who prioritize the development of MLSE can support students in exploring mathematical concepts and dealing with challenges. In addition, preservice teachers' attitudes toward mathematics are an important indicator of their self-efficacy beliefs regarding mathematical literacy (Önal *et al.*, 2017; Yavuz *et al.*, 2013). It is important in mathematics education because it affects motivation and success. Individuals with high levels of MLSE can use different strategies to interpret mathematical information and overcome real-life challenges (Kurniawati & Mahmudi, 2019). In this way, they can solve complex problems and clearly express their mathematical ideas to others. MLSE also includes thinking and communication (Rojas & Benakli, 2020). (Genç & Erbas, 2019) categorized teachers' conceptions of mathematical literacy into seven groups. MT is among them. It has been associated with mathematical literacy because it includes process-based skills such as abstraction, critical thinking, meaning, and prediction. Therefore, MLSE is also related to mathematical language and thinking skills. MT is a cognitive process that involves reasoning, problem-solving, and making sense of mathematical concepts and relationships. It goes beyond memorizing procedures and algorithms (Wu & Yang, 2022). MT focuses on understanding mathematical ideas, patterns, and structures. It utilizes various skills such as critical thinking, problem-solving, creativity, and metacognition. Thinking skills are also related to the perception of self-efficacy (Şahal *et al.*, 2021). Güneş and Gökçek (2013) reported that preservice teachers' literacy self-efficacy levels were above average, but preservice teachers perceived themselves as inadequate in mathematical thinking. While MT is necessary for success in mathematics education, it can be used academically and in the real world. Therefore, teachers, who have an important role in developing MT skills, must provide opportunities for students to explore mathematical concepts and improve their problem-solving abilities. This requires a high level of mathematical language self-efficacy. It is the belief in effectively understanding, using, and teaching mathematical language. It also includes using language effectively to express mathematical ideas, concepts, and relationships (Barçın & Yenmez, 2023). In addition, mathematical literacy self-efficacy perception and language use in mathematics teaching are also related to each other (Karademir & Deveci, 2019). Furthermore, mathematical language, which has communication and representation components, is one of the basic components in the development of mathematical literacy (Thompson & Chappell, 2007). Teachers with strong mathematical language self-efficacy can better explain

mathematical concepts. They create opportunities for students to engage with mathematical ideas individually and collaboratively. Thus, they encourage students to explain their mathematical ideas, justify their reasoning, and engage in mathematical discussions. In this way, students can increase their comprehension, problem-solving skills, and proficiency in mathematics.

### 1.3. Purpose of the Study

In recent years, many studies have been conducted to assess the professional mathematics competencies of pre-service teachers, developing different scales and analyzing the relationships between various variables. In particular, key variables such as mathematical literacy (Busnawir *et al.*, 2021; Topbaş Tat, 2018; Yavuz *et al.*, 2013), mathematical thinking skills (Ersoy & Başer, 2013; Wu & Yang, 2022; Yildiz *et al.*, 2020) and teacher self-efficacy (Berg *et al.*, 2024; Kabael & Yayan, 2017) were addressed within the scope of pre-service teachers' teaching competences. However, most of these studies use variable-centred analysis techniques and do not take into account the different competence patterns of individuals and their reflections on teacher education. Therefore, it is critical to identify the profiles of pre-service teachers' professional competencies in order to restructure the content, methods and goals of teacher education programmes. In this context, revealing how pre-service teachers construct latent profiles in the three main competence areas of MLSE, MT and TSEML fills an important gap both theoretically and practically. This study goes beyond traditional analyses to explain how these competencies are patterned together and aims to make an original contribution to the literature by using LPA, a person-centred analysis approach.

The originality of the study is that it deals with the constructs of MLSE, MT and TSEML, which are often analyzed independently of each other, from a holistic perspective. Whether pre-service teachers have different profiles according to their combinations in these three competence areas is still in need of research. Profile analysis studies of this kind will enable the development of more personalized and needs-based teacher education approaches to reinforce preservice teachers' strengths and support their areas for improvement. In this context, the primary objective of the study is to reveal the implicit profiles of pre-service teachers based on their MLSE, MT and TSEML levels and to discuss the meaning of these profiles for teacher education programmes. In addition, by describing the structural features of the profiles, it is aimed to develop a deeper understanding of pre-service teachers' professional development processes. In this respect, the study aims to provide original findings that can guide measurement-evaluation and individualized instruction practices in the field of teacher education.

## 2. METHOD

### 2.1. Research Design

The predictive correlational research model, one of the quantitative research approaches, was used (Şata, 2020). This model was preferred because it aims to describe the existing situation and examine the relationship between research variables.

### 2.2. Population-Sample and Power Analysis

A power analysis was conducted before the data collection process to determine the generalizability of the study results and the test's power. The conditions determined for the power analysis were alpha levels .05, beta level (type-II error) .20, small effect size, and two-way hypothesis. The minimum number for these conditions has been determined to be 384. The data collected was more than the minimum number, and post hoc power analysis was performed to determine the accuracy and strength of the results. As a result of the analysis, it was determined that the power of the test was  $(1 - \beta) = .936$ . According to the result, the analyses on the sample obtained have sufficient power.

The population is preservice teachers in the elementary mathematics education program in Turkey. The sample consists of 577 preservice teachers selected by convenient sampling method. As a result of missing data and outlier analysis, 32 ineligible participants were excluded from the study, and the research was conducted on 545 participants (69.9% female). Of the sample who continued their education at 4 different universities, 28.3% are at the 1st-grade level, 35.2% are at the 2nd-grade level, 19.4% are at the 3rd-grade level, and 17.1% are at the 4th-grade level. Selecting participants from different geographical regions and universities of Turkey and from different grade levels contributes to the generalizability of the research.

### 2.3. Data Collection

The data of the present study were collected through an online survey methodology. The questionnaire was created using Google Forms. All participants were informed that their participation in the study was voluntary and that the aims and procedures of the study were met. They were also asked for confirmation before completing the questionnaires. Confidentiality and anonymity of responses were assured for all participants. Participation in the study was entirely voluntary.

**2.3.1.1. Self-Efficacy Scale of Mathematics Literacy (SSML).** SSML developed by Özgen and Bindak (2008) consisting of 25 items, was designed to measure the self-efficacy perceptions regarding mathematical literacy of preservice teachers. Items (e.g., I can see mathematical relationships in scientific events) are scored on a five-point Likert scale from 1 (completely disagree) to 5 (completely agree). The score obtained from the scale varies between 25 and 125. High scores mean that preservice teachers' self-efficacy perceptions regarding mathematics literacy increase, while low scores mean their self-efficacy perceptions decrease. In the present study, the Cronbach  $\alpha$  coefficient calculated for the whole scale was .931, while the McDonald  $\omega$  coefficient was determined to be .933. CFA was performed for the validity of the measurements obtained from the measurement tool and the fit values were  $\chi^2 / df = 582.67 / 275 = 2.12$ , CFI = .981, NNFI = .980, NFI = .965, RMSEA (%90 CI) = .045 (.040 - .050), SRMR = .060 (Browne & Cudeck, 1999; Hu & Bentler, 1999; Kline, 2015).

**2.3.1.2. Mathematical Thinking Scale (MTS).** MTS, developed by Ersoy and Başer (2013), consisting of 25 items, was designed to measure the mathematical thinking levels of preservice teachers. Items (e.g., An individual who has creative thinking skills acquires mathematical thinking skills more easily) are scored on a five-point Likert scale from 1 (completely disagree) to 5 (completely agree). The score obtained from the scale varies between 25 and 125. High scores mean preservice teachers' mathematical thinking levels increase, while low scores mean their thinking level decreases. MTS consists of four subscales. These are high-level thinking tendencies, reasoning, mathematical thinking, and problem-solving. In the present study, it was determined that the Cronbach  $\alpha$  coefficient calculated for the subscales was between .591 - .931, while the McDonald  $\omega$  coefficient varied between .621 - .933, and the stratified Cronbach  $\alpha$  coefficient calculated for the whole scale was .897. CFA was performed for the validity of the measurements obtained from the measurement tool, and the fit values were  $\chi^2 / df = 648.95 / 269 = 2.41$ , CFI = .968, NNFI = .964, NFI = .947, RMSEA (%90 CI) = .051 (.046 - .056), SRMR = .066.

**2.3.1.3. Teacher Self-Efficacy Scale in Language of Mathematics (TSESLoM).** TSESLoM, developed by Kabael and Yayan (2017), consisting of 17 items, was designed to measure the self-efficacy perception levels of preservice teachers using and teaching the language of mathematics. Items (e.g., I can use mathematical language to express mathematical thoughts) are scored on a four-point Likert scale from 1 (completely disagree) to 4 (completely agree). The score obtained from the scale varies between 17 and 68. High scores mean that preservice teachers' self-efficacy perception levels for using and teaching mathematical language increase, while low scores mean that self-efficacy perception level for using and teaching mathematical language decreases. TSESLoM consists of three subscales. These are



teaching the language of mathematics, using the specific language of mathematics, and using the general language of mathematics. In the present study, it was determined that the Cronbach  $\alpha$  coefficient calculated for the subscales was between .740 - .796, while the McDonald  $\omega$  coefficient varied between .752 - .818, and the stratified Cronbach  $\alpha$  coefficient calculated for the whole scale was .876. CFA was performed for the validity of the measurements obtained from the measurement tool, and the fit values were  $\chi^2/df = 486.24 / 116 = 4.19$ , CFI = .952, NNFI = .944, NFI = .938, RMSEA (%90 CI) = .077 (.070 - .084), SRMR = .082.

## 2.4. Data Analysis

Descriptive statistics of the measurements, Cronbach alpha and McDonald omega coefficients for the reliability of the measurements, and confirmatory factor analysis (CFA) were performed to provide evidence for the validity of the measurements. In this study, the threshold values of the fit indices used in confirmatory factor analysis (CFA) were determined based on standards accepted in the relevant literature. In particular, a value of  $\geq .90$  for CFI (Comparative Fit Index) and TLI (Tucker-Lewis Index) indicates that the model provides a good level of fit (Hu & Bentler, 1999; Kline, 2015). These indices were preferred because they are sensitive in measuring poor fit compared to the model and because values above .90 increase the explanatory power of the model. On the other hand, RMSEA (Root Mean Square Error of Approximation) and SRMR (Standardized Root Mean Square Residual) values  $\leq .08$  indicate that the error rate of the model is within reasonable limits and that it provides an adequate level of fit with the data (Hu & Bentler, 1999; Kline, 2015). Although the RMSEA value is affected by sample size, values below .08 are generally considered as “acceptable fit” criteria. For SRMR, which reflects the difference between the observed and expected correlations in the model, values below .08 indicate that the differences between the data and the model are low. These threshold values were chosen to ensure that the findings of the study are comparable to international standards, to reduce the risk of overfitting the model, and to support the reliability of the research. LPA was conducted to determine the profiles of preservice elementary mathematics teachers. Jamovi, JASP, SPSS, and Mplus package programs were used in data analysis. In data analysis, the .05 level was considered for statistical significance.

## 3. RESULTS

Descriptive statistics calculated for the measurements obtained from the measurement tools are presented in Table 1.

**Table 1.** Descriptive statistics of measurements.

Scales (Variables)	Min	Max	$\bar{X}$	SD	Skewness	SE	Kurtosis	SE
SSML (MLSE)	48	124	91.13	14.13	0.063	0.105	-0.066	0.209
MTS1 (MT-HLTD)	8	20	16.61	2.50	-0.462	0.105	-0.397	0.209
MTS2 (MT-R)	13	30	24.51	3.61	-0.443	0.105	-0.177	0.209
MTS3 (MT-MTS)	17	40	29.86	3.69	-0.193	0.105	-0.020	0.209
MTS4 (MT-PS)	16	35	26.04	3.32	-0.073	0.105	-0.263	0.209
TSESLoM1 (TSEML-TLoM)	4	16	12.15	2.27	-0.030	0.105	-0.329	0.209
TSESLoM2(TSEML-USLoM)	7	20	16.20	2.61	-0.330	0.105	-0.191	0.209
TSESLoM3(TSEML-UGLoM)	13	32	24.04	3.83	-0.011	0.105	-0.764	0.209

HLTD: High-level thinking tendency, R: Reasoning, MTS: Mathematical thinking skill, PS: Problem solving, TLoM: Teaching language of mathematics, USLoM: Using specific language of mathematics, UGLoM: Using general language of mathematics

Table 1 shows that individuals' averages for all measurement tools are high. Skewness and kurtosis values of the measurements are in the range of  $\pm 1,000$ . Accordingly, it was determined that the measurements had a normal distribution (Shiel & Cartwright, 2015). Pearson correlation coefficient was used to determine the relationships between measurement tools, and

the findings are given in Table 2. Table 2 shows that the variables are positively correlated at low and medium levels ( $p < .05$ ). The significance of relationships between variables, an assumption of profile analysis, appears to have been met.

Latent profiles of preservice teachers were created using mathematical literacy self-efficacy, sub-dimensions of mathematical thinking (HLTD, R, MTS, and PS), and sub-dimensions of using mathematical language (TLoM, USLoM, and UGLoM). LPA was used to determine the profiles. In LPA, AIC, BIC, corrected BIC (D-BIC), Lo-Mendell-Rubin likelihood ratio test (LMR-LRT), and entropy coefficient were used to decide the number of profiles (Spurk *et al.*, 2020).

**Table 2.** Bivariate correlations between variables.

Variables	1	2	3	4	5	6	7
1. MLSE	--						
2. MT- HLTD	.445**	--					
3. MT-R	.495**	.742**	--				
4. MT-MTS	.366**	.576**	.623**	--			
5. MT- PS	.476**	.529**	.611**	.490**	--		
6. TSEML-TLoM	.620**	.317**	.338**	.280**	.350**	--	
7. TSEML-USLoM	.598**	.313**	.344**	.295**	.356**	.613**	--
8. TSEML-UGLoM	.553**	.388**	.407**	.444**	.430**	.518**	.483**

\*\* $p < .001$ .

LPA has a stepwise process, and initially, two profiles are estimated. Then, the number of profiles is increased and continued. According to the fit values, the best profile number is determined, and the analysis is ended. In the present study, analyses ranging from 2 to 6 profiles were made, and the results are presented in Table 3.

**Table 3.** Model fit statistics for determining the optimal number of profiles.

Model	AIC	BIC	ABIC	$p$ -values of LMR test	$p$ -values of BLRT test	Entropy
Profile 2	22598.57	22706.09	22626.73	.000	.000	.839
<b>Profile 3</b>	<b>22162.51</b>	<b>22308.74</b>	<b>22200.81</b>	<b>.000</b>	<b>.000</b>	<b>.851</b>
Profile 4	21983.23	22168.16	22031.66	.066	.000	.821
Profile 5	21891.48	22115.12	21950.05	.067	.000	.815
Profile 6	21834.71	22097.06	21903.42	.275	.000	.831

Note. Optimal number in bold.

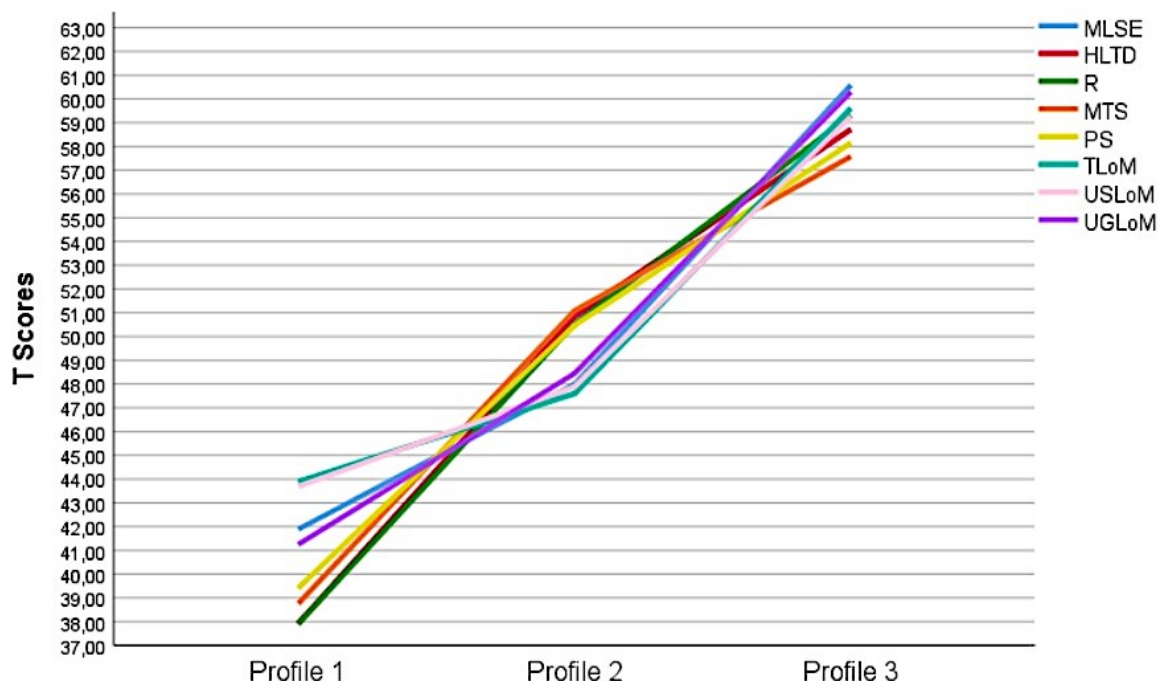
Table 3 shows that the estimated values for 3 profiles are appropriate according to the criteria for determining the optimal number of profiles. It is seen that the entropy value starts to decrease after 3 profiles. An entropy value close to 1.00 indicates that the model is better. This situation is in favor of the 3-profile solution. Moreover, since BIC and AIC values are very close, using a model with 3 profiles (Saatçioğlu, 2023; Spurk *et al.*, 2020). As a result, 3 profile models were selected, and their differences according to classes were examined. The graphical representation of the 3-profile model is presented in Figure 1. Figure 1 shows preservice teachers' standardized averages (T scores) in each profile. The LPA results revealed three different profiles: Profile 1 (Low Competence Group), Profile 2 (Moderate Competence Group) and Profile 3 (High Competence Group). Each profile showed significant differences in mathematical literacy, mathematical thinking and mathematical language self-efficacy.

The pre-service teachers in Profile 1 had the lowest T scores in all scales (in the range of 40-45). Low mathematical literacy indicates that they have difficulty in applying basic

mathematical concepts in the context of daily life. Low mathematical thinking indicates that students have inadequate problem solving, reasoning and relationship building skills. Furthermore, poor mathematical language self-efficacy indicates that these students have low confidence in expressing mathematical ideas in verbal or writing.

The pre-service teachers in Profile 2 had a moderate level of competence (T scores in the range of approximately 47-52). Pre-service teachers in this profile are able to understand basic mathematical concepts and have moderate mathematical thinking skills. Their self-efficacy in mathematical language is higher compared to Profile 1 and lower compared to Profile 3. These findings indicate that pre-service teachers in Profile 2 have a moderate level of self-confidence in performing mathematical tasks, but may have difficulties with complex mathematical problems.

**Figure 1.** Graphical representation comparing profiles of the variables in T-score format.



On the other hand, pre-service teachers in Profile 3 have a high level of competence (T scores in the range of 55-62). They exhibit high achievement in mathematical literacy and mathematical thinking and have high self-efficacy perception in using mathematical language. In other words, they are able to use mathematical concepts effectively in daily life and produce creative and analytical solutions by comprehending abstract mathematical relationships. They also have high confidence in their ability to explain and communicate mathematical expressions.

These findings support an important point emphasized in the literature regarding the use of LPA. Latent profiles reveal not only quantitative differences between groups but also qualitative variations in individuals' competence domains (Collins & Lanza, 2010). The results of the study show how mathematical literacy, mathematical thinking and mathematical language self-efficacy constructs play a role together in the professional development of pre-service teachers and emphasize the importance of designing educational intervention programmes to strengthen these three competence domains together. It is stated that it is important that the LPA not only identifies the groups but also shows how these groups are related to the basic concepts and variables in the research (Berlin *et al.*, 2014; Pastor *et al.*, 2007). In this respect, each of the profiles identified in the study was comprehensively defined, and how each profile was associated with the three main competence areas examined in the study was clearly demonstrated.

#### 4. DISCUSSION and CONCLUSION

In the present study, the latent model of MLSE, MT, and TSEML was examined to determine the mathematical competencies of preservice teachers. This aimed to identify the strengths and weaknesses of preservice teachers during the training period for teaching mathematics. The analysis results show that the three-profile solution best fits the model. These profiles provide important information about preservice teachers' mathematical competence levels. Preservice teachers in Profile 1 exhibited below-average levels of mathematical competence. In particular, the low level of MLSE and TSEML skills in this profile implies that they may have difficulty teaching mathematical concepts to students (Ball *et al.*, 2008). The low performance of preservice teachers in Profile 1 in TSEML and MLSE overlaps with the deficiencies in pedagogical content knowledge identified in the MKT model (Ball *et al.*, 2008). This indicates that these preservice teachers may struggle to effectively transfer mathematical concepts to students (Yildiz & Arpaci, 2024). Moreover, the level of MT was lower than the other variables. This may be associated with the lack of systematic thinking in mathematical problem-solving processes (Uras & Soylu, 2024). In Profile 2, all variables are at an average level, but unlike Profile 1, MT is at the highest level. This profile indicates that MT skills can develop independently of other competencies. However, the average level of TSEML and MLSE skills indicates that preservice teachers may not be able to effectively integrate mathematical knowledge in a pedagogical context (Blömeke *et al.*, 2020). Finally, Profile 3 is the profile where all variables are above the average, and the variation between variables is the least. This profile can be interpreted as an ideal preservice teacher profile in which mathematical competencies are developed in a balanced way. In particular, the strong correlation between TSEML and MLSE supports that mathematical language skills are directly related to literacy (Zhang & Tian, 2024).

The findings of this study clearly reveal that there is not a homogeneous structure in pre-service teachers' mathematical competences. In particular, the different levels of competences in each of MLSE, MT and TSEML skills indicate that pre-service teachers need to be supported individually in their education. This indicates that standardized teacher education practices may not be effective for all pre-service teachers (Blömeke *et al.*, 2020; Yang *et al.*, 2020). For example, while TSEML and MLSE-focused interventions can be designed for preservice teachers in Profile 1, preservice teachers in Profile 2 can benefit from trainings that integrate MT skills with pedagogical practices. In particular, it is consistent with the literature that low performance in MLSE and TSEML skills is associated with preservice teachers' lack of pedagogical content knowledge (Ball *et al.*, 2008). These differences suggest that approaches that focus on individual needs should be adopted in teacher education (Blömeke *et al.*, 2020). Moreover, as Yang *et al.* (2020) emphasized, this heterogeneous structure supports the need for teacher education policies to be designed flexibly according to preservice teachers' initial qualification levels.

Examining the relationships between variables is an important step in mathematics competencies. The correlation between MT and TSEML (Profile 3) indicates that mathematical thinking skills are supported by language use. This aligns with Sociocultural Theory (Steele, 2001) and emphasizes the importance of interdisciplinary integration in teacher education. This emphasizes the need to approach these competencies in teacher education in an integrated way rather than separately (Zhang & Tian, 2024). Low to moderate positive correlations indicate that mathematical literacy, thinking, and language skills are interrelated. In addition, the fact that the level of mathematical competencies varies similarly in all profiles shows that professional competencies may be related to each other (Buchholtz *et al.*, 2023). This highlights the importance of designing teacher education programs with a holistic approach. It also indicates that dealing with various mathematical skills of preservice teachers together can increase the effectiveness of education (Olawale, 2024). It has been observed that as mathematical thinking skills improve, their ability to use the mathematical language also



increases. This finding reveals the importance of strategies that strengthen mathematical thinking and language skills in teacher education. The findings show that pre-service teachers' mathematical competencies are shaped by the interaction of skills. Furthermore, this study makes important contributions to the international literature. First, the necessity of a multidimensional approach to teacher competencies is consistent with the literature based on individual-centered analyses (Blömeke *et al.*, 2020; Schel & Drechsel, 2025). Secondly, the possibility of categorizing pre-service teachers' competence constructs in more detail enables the development of appropriate intervention strategies for the profile.

Identifying pre-service teachers' different mathematical competence profiles through LPA makes an important contribution to teacher education and development. LPA reveals not only pre-service teachers' performances in individual variables but also the holistic structures formed by multiple competence areas such as mathematical literacy, thinking and teaching self-efficacy (Blömeke *et al.*, 2022; Vilppu *et al.*, 2024). The competence levels of pre-service teachers form important profiles among pre-service teachers and the homogeneity levels of the profiles differ. In particular, the low performance of the pre-service teachers in the first profile in all competency areas shows the necessity of targeted support and development strategies for specific groups. This indicates that teacher education programs need to be structured in a flexible and tailored manner according to the competence combinations of pre-service teachers. Thus, LPA can be considered not only as a diagnostic tool but also as an implementing tool that guides teacher education policies shaped according to individual competency profiles. In conclusion, this study reveals a model used to determine the mathematics proficiency profiles of preservice teachers. Identifying different mathematics proficiency profiles is important in teacher training and development. These profiles can provide a basis for supporting preservice teachers based on their strengths and weaknesses.

## 5. LIMITATIONS and RECOMMENDATIONS

This study has several limitations. First, the study focused only on the mathematics competencies of preservice teachers and did not include the competencies of in-service teachers. This may make it difficult to generalize the findings to the general teacher population. Secondly, the research findings are based only on the LPA method. This may be a limitation in understanding preservice teachers' mathematical competencies in depth. Finally, the analysis was limited to only three components (mathematical literacy self-efficacy, mathematical thinking and teacher efficacy in mathematical language). It means that other important elements that may influence teacher efficacy (e.g. problem solving, logical reasoning, pedagogical competences) were left out.

Some suggestions can be made for future research. First, examining the mathematical competence profiles of preservice and in-service teachers may provide more comprehensive information about teacher education and professional development processes. Second, mixed-methods studies supported by qualitative research can help to gain a deeper understanding of preservice teachers' mathematical competencies. Such studies can provide more detailed insights into how preservice teachers develop their competencies. Finally, assessing other components of mathematical competencies, such as problem-solving and logical thinking, can help preservice teachers develop a more holistic approach to mathematics education. In addition, longitudinal studies must be conducted to investigate the changes in mathematical competencies over time and their relationship with additional variables such as problem-solving. Restructuring education programs to consider this diversity can contribute to teachers' professional development.

## Acknowledgments

This study was not supported by any funding or grant. The summary of this study was presented as an oral presentation at the TRB2 International Congress on Educational Sciences-II (October 20–21, 2023).

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Ağrı İbrahim Çeçen University Ethics Committee, 30.05.2024, 200.

### Contribution of Authors

**Muhammed Celal Uras:** Investigation, Resources, Visualization, Software, Formal analysis and Writing-original draft. **Mehmet Şata:** Methodology and Validation. **Yasin Soylu:** Literature review and Supervision.

### Orcid

Muhammed Celal Uras  <https://orcid.org/0000-0003-3994-8723>

Mehmet Şata  <https://orcid.org/0000-0003-2683-4997>

Yasin Soylu  <https://orcid.org/0000-0003-0906-4994>

### REFERENCES

- Akkaya, R., Sezgin Memnun, D., & Katranci, Y. (2012, March 5-9). *Teacher trainees' self-efficacy beliefs about mathematical literacy: Turkey case* [Paper presentation]. 23<sup>rd</sup> International Conference Society for Information Technology and Teacher Education, Austin, Texas, United States.
- Aktulun, Ö.U. (2018). Examination of the relationships between mathematics literacy self-efficacy perceptions of preschool teachers and geometric shape recognition and number skills of children with structural equation modelling. *International Education Studies*, 11(12), 63-77. <https://doi.org/10.5539/ies.v11n12p63>
- Ball, D.L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389-407. <https://doi.org/10.1177/0022487108324554>
- Barçın, N., & Yenmez, A.A. (2023). Geogebra software on the mathematical language developments and self-efficacy perceptions of students. *International e-Journal of Educational Studies*, 7(15), 682-704. <https://doi.org/10.31458/iejes.1340349>
- Berg, D.A.G., Ingram, N., Asil, M., Ward, J., & Smith, K.J. (2025). Self-efficacy in teaching mathematics and the use of effective pedagogical practices in New Zealand primary schools. *Journal of Mathematics Teacher Education*, 28, 129-149. <https://doi.org/10.1007/s10857-024-09623-9>
- Berlin, K.S., Williams, N.A., & Parra, G.R. (2014). An introduction to latent variable mixture modeling (part 2): Longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology*, 39(2), 188–203. <https://doi.org/10.1093/jpepsy/jst085>
- Blömeke, S., Kaiser, G., König, J., & Jentsch, A. (2020). Profiles of mathematics teachers' competence and their relation to instructional quality. *ZDM Mathematics Education*, 52, 329-342. <https://doi.org/10.1007/s11858-020-01128-y>
- Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, 79, Article 101600. <https://doi.org/10.1016/j.learninstruc.2022.101600>
- Browne, M.W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230-258. <https://doi.org/10.1177/0049124192021002005>
- Buchholtz, N., Kaiser, G., & Schwarz, B. (2023). The evolution of research on mathematics teachers' competencies, knowledge and skills. In A. Manizade, N. Buchholtz, & K. Beswick (Eds.), *The evolution of research on teaching mathematics: International perspectives in the digital Era* (pp. 55-89). Springer. [https://doi.org/10.1007/978-3-031-31193-2\\_3](https://doi.org/10.1007/978-3-031-31193-2_3)

- Busnawir, B., La, M., Sudia, M., Idris, M., & Sadikin, S. (2021). Analysis of mathematical literacy ability in terms of self-efficacy high and low. *International Journal of New Trends in Arts, Sports & Science Education*, 10(5), 316-325.
- Collins, L.M., & Lanza, S.T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Wiley. <https://doi.org/10.1002/9780470567333>
- Cresswell, C., & Speelman, C.P. (2020). Does mathematics training lead to better logical thinking and reasoning? A cross-sectional assessment from students to professors. *PLoS One*, 15(7), e0236153. <https://doi.org/10.1371/journal.pone.0236153>
- Dervenis, C., Fitsilis, P., & Iatrellis, O. (2022). A review of research on teacher competencies in higher education. *Quality Assurance in Education*, 30(2), 199-220. <https://doi.org/10.1108/QAE-08-2021-0126>
- Ekmekci, A., & Serrano, D.M. (2022). The impact of teacher quality on student motivation, achievement, and persistence in science and mathematics. *Education Sciences*, 12(10), 649. <https://doi.org/10.3390/educsci12100649>
- Ersoy, E., & Başer, N. (2013). Matematiksel düşünme ölçeğinin geliştirilmesi [The development of mathematical thinking scale]. *Kastamonu Education Journal*, 21(4), 1471-1486. <https://dergipark.org.tr/en/pub/kefdergi/issue/22604/241557>
- Gardesten, M. (2023). How co-teaching may contribute to inclusion in mathematics education: A systematic literature review. *Education Sciences*, 13(7), 677. <https://doi.org/10.3390/educsci13070677>
- Genç, M., & Erbas, A.K. (2019). Secondary mathematics teachers' conceptions of mathematical literacy. *International Journal of Education in Mathematics, Science and Technology*, 7, 222-237.
- Güneş, G., & Gökçek, T. (2013). Öğretmen adaylarının matematik okuryazarlık düzeylerinin belirlenmesi [Determination of preservice teachers' mathematical literacy levels]. *Dicle University Journal of Ziya Gökalp Education Faculty*, 20, 70-79.
- Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 1, 1–55. <https://doi.org/10.1080/10705519909540118>
- Jentsch, A., & König, J. (2022). Teacher competence and professional development. In T. Nilsen, A. Stancel-Piatak, & J.-E. Gustafsson (Eds.), *International handbook of comparative large-scale studies in education: Perspectives, methods and findings* (pp. 1167-1183). Springer. [https://doi.org/10.1007/978-3-030-88178-8\\_38](https://doi.org/10.1007/978-3-030-88178-8_38)
- Kabael, T., & Yayan, B. (2017). Effect of self-evaluation on preservice mathematics teachers' self-efficacy in the language of mathematics. *Anadolu Journal of Educational Sciences International*, 7(1), 1-34.
- Karademir, Ç.A., & Deveci, Ö. (2019). Sınıf öğretmeni adaylarının matematik öğretiminde matematik dili kullanımları ve matematik okuryazarlığı öz yeterlik algıları [Primary pre-service teachers' mathematical language usage in mathematics instruction and mathematical literacy self-efficacy perceptions]. *İnönü University Journal of the Faculty of Education*, 20(3), 695-708. <https://doi.org/10.17679/inuefd.419755>
- Kline, R.B. (2015). *Principles and practice of structural equation modeling* (4<sup>th</sup> ed.). The Guilford Press.
- Kurniawati, N.D.L., & Mahmudi, A. (2019). Analysis of mathematical literacy skills and mathematics self-efficacy of junior high school students. *Journal of Physics: Conference Series*, 1320(1), Article 012053. <https://doi.org/10.1088/1742-6596/1320/1/012053>

- Mukuka, A., Balimuttajjo, S., & Mutarutinya, V. (2023). Teacher efforts towards the development of students' mathematical reasoning skills. *Heliyon*, 9(4), Article e14789. <https://doi.org/10.1016/j.heliyon.2023.e14789>
- Nilimaa, J. (2023). New examination approach for real-world creativity and problem-solving skills in mathematics. *Trends in Higher Education*, 2(3), 477-495. <https://doi.org/10.3390/higheredu2030028>
- OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. <https://doi.org/10.1787/7ea9ee19-en>
- Olawale, B.E. (2024). Impact of preservice teacher education programme on mathematics student teachers' teaching practices during school experiences. *Education Sciences*, 14(7), Article 762. <https://doi.org/10.3390/educsci14070762>
- Ozgen, K. (2019). Problem-posing skills for mathematical literacy: The sample of teachers and preservice teachers. *Eurasian Journal of Educational Research*, 19(84), 179-212. <https://doi.org/10.14689/ejer.2019.84.9>
- Önal, H., Yorulmaz, A., Gökbulut, Y., & Çilingir Altiner, E. (2017). The relationship between preservice class teachers' self-efficacy in mathematical literacy and their attitudes towards mathematics. *Journal of Education and Practice*, 8, 170-179.
- Özgen, K., & Bindak, R. (2008). Matematik okuryazarlığı öz-yeterlik ölçeğinin geliştirilmesi [The development of self-efficacy scale for mathematics literacy]. *Kastamonu Education Journal*, 16(2), 517-528.
- Pastor, D.A., Barron, K.E., Miller, B.J., & Davis, S.L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, 32(1), 8-47. <https://doi.org/10.1016/j.cedpsych.2006.10.003>
- Peter, E.E. (2012). Critical thinking: Essence for teaching mathematics and mathematics problem solving skills. *African Journal of Mathematics and Computer Science Research*, 5(3), 39-43. <https://doi.org/10.5897/AJMCSR11.161>
- Rojas, E., & Benakli, N. (2020). Mathematical literacy and critical thinking. In J. C. But (Ed.), *Teaching college-level disciplinary literacy: Strategies and practices in STEM and professional studies* (pp. 197-226). Springer. [https://doi.org/10.1007/978-3-030-39804-0\\_8](https://doi.org/10.1007/978-3-030-39804-0_8)
- Saatçioğlu, F.M. (2023). Örtük profil analizi ile öğrencilerin matematik tutum profillerinin belirlenmesi üzerine bir araştırma [A study on determining students' mathematical attitude profiles by latent profile analysis]. *Gazi University Gazi Faculty of Education Journal*, 43(3), 1623-1643. <https://doi.org/10.17152/gefad.1352037>
- Schel, J., & Drechsel, B. (2025). A latent profile analysis for teacher education students' learning: An overview of competencies in self-regulated learning. *Frontiers in Psychology*, 16, Article 1527438. <https://doi.org/10.3389/fpsyg.2025.1527438>
- Schwarz, B., & Kaiser, G. (2019). The professional development of mathematics teachers. In G. Kaiser & N. Presmeg (Eds.), *Compendium for early career researchers in mathematics education* (pp. 325-343). Springer. [https://doi.org/10.1007/978-3-030-15636-7\\_15](https://doi.org/10.1007/978-3-030-15636-7_15)
- Shiel, G., & Cartwright, F. (2015). *Analyzing Data from a National Assessment of Educational Achievement* (Vol. 4). World Bank Publications.
- Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and “how to” guide of its application within vocational behavior research. *Journal of Vocational Behavior*, 120, Article 103445. <https://doi.org/10.1016/j.jvb.2020.103445>
- Steele, D.F. (2001). Using sociocultural theory to teach mathematics: A Vygotskian perspective. *School Science and Mathematics*, 101(8), 404-416. <https://doi.org/10.1111/j.1949-8594.2001.tb17876.x>
- Şahal, M., Özdemir, A., & Karaşan, S. (2021). An examination of the relationship between secondary school students' abstract thinking skills, self-efficacy perceptions and attitudes



- towards mathematics. *Participatory Educational Research*, 8(2), 391-406. <https://doi.org/10.17275/per.21.45.8.2>
- Şata, M. (2020). Nicel araştırma yaklaşımları [Quantitative research approaches] In E. Oğuz (Ed.), *Eğitimde Araştırma Yöntemleri* (pp. 77-98). Eğiten Kitap.
- Thompson, D.R., & Chappell, M.F. (2007). Communication and representation as elements in mathematical literacy. *Reading & Writing Quarterly*, 23(2), 179-196. <https://doi.org/10.1080/10573560601158495>
- Topbaş Tat, E. (2018). Prospective mathematics teachers' perceived self-efficacy in mathematical literacy. *Elementary Education Online*, 17(2), 489-499. <https://doi.org/10.17051/ilkonline.2018.418887>
- Uras, M.C., Şata, M., & Soylu, Y. (2024). Investigation of preservice teachers' statistical literacy levels. *International Journal of Educational Studies and Policy*, 5(2), 174-185. <https://doi.org/10.5281/zenodo.14016307>
- Uras, M.C., & Soylu, Y. (2024). Investigation of elementary school mathematics teachers' problem-solving skills without variables and problem-solving strategies. *Technology, Innovation and Special Education Research*, 4(2), 168-191.
- Uras, M.C., Kaya, S., Kaya, A., & Yildirim, M. (2025). Unveiling risk profiles: A latent profile analysis of 21st-century skills, resistance to change, and cognitive flexibility. *Brain and Behavior*, 15(1), e70167. <https://doi.org/10.1002/brb3.70167>
- Vilppu, H., Laakkonen, E., Laine, A., Lähteenmäki, M., Metsäpelto, R.L., Mikkilä-Erdmann, M., & Warinowski, A. (2024). Learning strategies, self-efficacy beliefs and academic achievement of first-year preservice teachers: A person-centred approach. *European Journal of Psychology of Education*, 39, 1161–1186. <https://doi.org/10.1007/s10212-023-00729-x>
- Wu, W.-R., & Yang, K.-L. (2022). The relationships between computational and mathematical thinking: A review study on tasks. *Cogent Education*, 9(1), Article 2098929. <https://doi.org/10.1080/2331186X.2022.2098929>
- Yang, X., Kaiser, G., König, J., & Blömeke, S. (2020). Relationship between preservice mathematics teachers' knowledge, beliefs and instructional practices in China. *ZDM*, 52(2), 281-294. <https://doi.org/10.1007/s11858-020-01145-x>
- Yavuz, G., Gunhan, B.C., Ersoy, E., & Narli, S. (2013). Self-efficacy beliefs of prospective primary mathematics teachers about mathematical literacy. *Journal of College Teaching & Learning*, 10(4), 279-288. <https://doi.org/10.19030/tlc.v10i4.8124>
- Yildiz, A., Baltacı, S., & Kartal, B. (2020). Examining the relationship between preservice mathematics teachers' mathematical thinking level and attitude towards mathematics courses. *Acta Didactica Napocensia*, 13(2), 256-270. <https://doi.org/10.24193/adn.13.2.17>
- Yildiz, E., & Arpacı, I. (2024). Understanding preservice mathematics teachers' intentions to use GeoGebra: The role of technological pedagogical content knowledge. *Education and Information Technologies*, 29(14), 18817-18838. <https://doi.org/10.1007/s10639-024-12614-1>
- Zehir, K., & Zehir, H. (2016). İlköğretim matematik öğretmen adaylarının matematik okuryazarlığı öz-yeterlik inanç düzeylerinin çeşitli değişkenler açısından incelenmesi [Investigation of elementary mathematics student teachers' mathematics literacy self-efficacy beliefs according to some variables]. *International Journal of Education, Science and Technology*, 2(2), 104-117.
- Zhang, H., & Tian, M. (2024). Unpacking the multi-dimensional nature of teacher competencies: A systematic review. *Scandinavian Journal of Educational Research*, 69(5), 1004–1025. <https://doi.org/10.1080/00313831.2024.2369867>



## Answer-based and reference-based BERT models for automatic scoring of Turkish short answers: The decisive role of task complexity

Abdulkadir Kara <sup>1\*</sup>, Zeynep Avinç Kara <sup>2</sup>, Serkan Yıldırım <sup>3</sup>

<sup>1</sup>Bayburt University, Department of Distance Education Application and Research Center, Bayburt, Türkiye

<sup>2</sup>TC Ministry of National Education, Erzurum, Türkiye

<sup>3</sup>Atatürk University, Kazım Karabekir Faculty of Education, Department of Computer Education and Instructional Technology, Erzurum, Türkiye

### ARTICLE HISTORY

Received: Apr. 30, 2025

Accepted: Aug. 28, 2025

### Keywords:

Artificial intelligence,  
Deep learning,  
Automatic scoring,  
Short answer,  
BERT LLM models.

**Abstract:** In measurement and evaluation processes, natural language responses are often avoided due to time, workload, and reliability concerns. However, the increasing popularity of automatic short-answer grading studies for natural language responses means such answers can now be measured more quickly and reliably. This study aims to build models for predicting automatic short answer scores using the pre-trained BERT deep learning language model and to reveal their effectiveness. For this purpose, two different score prediction models were created using an answer-based approach that aligns student answers with expert judgements and a reference-based approach that matches student answers with reference answers. The dataset includes answers from 246 Physics department students responding to 4 physics-related questions. The performance of these models was evaluated on four physics questions representing varying levels of cognitive complexity, using Cohen's Kappa for statistical comparison of agreement with expert scores. Our findings reveal a clear interaction between model architecture and task complexity. The answer-based model was unequivocally superior for the most complex, multi-class task, effectively capturing diverse, nuanced responses. Conversely, the reference-based model demonstrated a statistically significant advantage for a well-defined, medium-complexity binary task. This study concludes that the optimal model for ASAG in Turkish is contingent on the cognitive demands of the assessment task, suggesting that a onsize-fits-all solution may not be the most effective approach. This provides a critical framework for practitioners, demonstrating not only that effective models are feasible for complex languages, but that their selection must be guided by task complexity.

## 1. INTRODUCTION

Assessing what and how students learn has become increasingly critical in today's educational landscape, where instructional quality and accountability are closely intertwined. Measurement and evaluation are crucial in understanding educational effectiveness (Kurbanoğlu & Olcaytürk, 2023). Preferred understandings greatly influence students' learning outcomes (Yıldırım & Bilican-Demir, 2022). Choice-based techniques are commonly used in learning

\*CONTACT: Abdulkadir KARA ✉ [abdulkadirkara@bayburt.edu.tr](mailto:abdulkadirkara@bayburt.edu.tr) 📍 Bayburt University, Department of Distance Education Application and Research Center, Bayburt, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

environments (Benli & İsmailova, 2018; Katsaris & Vidakis, 2021). Nevertheless, such techniques involve students selecting an option without justifying it (Çınar *et al.*, 2020), leading to random choices and undermining the validity of scores. Moreover, it can be difficult to identify lasting learning outcomes with this method. These issues arising from choice-based techniques require implementing measurement and evaluation techniques in academic settings and diversifying strategies.

Several factors stand out when considering the frequent use of choice-based techniques. The evaluation process of Natural Language Responses (NLR) increases the workload for teachers (Uyar & Büyükahıska, 2025). As a result, the evaluation process takes longer (Westera *et al.*, 2018). It becomes inapplicable, especially in crowded learning environments (Chen *et al.*, 2025). Additionally, there is a potential for including subjective evaluations from teachers within the score results, which may overshadow the process (Abdul-Salam *et al.*, 2022). Subjective judgements pose a risk to the reliability of scoring in evaluations. NLRs are less favored in learning environments due to increased workload, time requirements, and reliability concerns. Choice-based techniques are preferable as they offer quick and dependable measurements (Garg *et al.*, 2022; Hasanah *et al.*, 2016). However, they exhibit a limited ability to recognize learning situations. There is insufficient evidence for the detection of deep and meaningful learning with choice-based techniques (Noyes *et al.*, 2020). Students' ability to make random choices makes it difficult to accurately measure their cognitive levels (Zhu *et al.*, 2022).

The limitations imposed by choice-based techniques for identifying learning situations have increased research on Natural Language Processing (NLP) and NLR (Burrows *et al.*, 2015). Especially, technological developments have motivated research in this field (Jadidinejad & Mahmoudi, 2014). Some studies have addressed issues related to the structure of NLR to solve or mitigate the related problems. String-based research on NLRs has focused on word pairings and predicted sentence similarities (Leacock & Chodorow, 2003; Siddiqi *et al.*, 2010). Semantic-based research, which focuses on the meaning of the responses, uses pre-trained word vectors such as GloVe, FastText, and Word2Vec to analyze words semantically (Lubis *et al.*, 2021; Saunders *et al.*, 2014; Zehner *et al.*, 2016). With the advances in technology, machine learning and deep learning-based research has come to the forefront and is more effective in complex language structures than other research structures (Gomaa *et al.*, 2023; Li *et al.*, 2022; Tulu *et al.*, 2021; Uysal & Dogan, 2021). The reason for these efforts is that the advantages that the smoothly functioning NLR provides to the measurement and evaluation processes are important. Using NLR provides essential evidence for rigorous evaluation of the learning process (Westera *et al.*, 2018) and for developing and improving learning approaches (Noyes *et al.*, 2020). NLR provides and assesses learners' ability to accurately recall and communicate information (Uto & Uchida, 2020).

The common point of these studies in the literature is that they focus on automatic scoring of NLRs. Automatic scoring of NLRs stands out due to consistent and objective grading, reduced human labor, and time-saving rapid evaluation processes (Abdul-Salam *et al.*, 2022; Dönmez, 2024; Uyar & Büyükahıska, 2025). Automatic scoring studies on natural language began with Page's (1967) work as a secondary school teacher in the 1960s (Ramineni & Williamson, 2013). With the advancement in technology and research in natural language processing, its popularity has increased since the 2000s. Since 2010, numerous studies have been conducted on grading NLR (Filighera *et al.*, 2023; Ghavidel *et al.*, 2020; Saunders *et al.*, 2014; Tulu *et al.*, 2021; Zimmerman *et al.*, 2018). The rise of online learning environments has sparked interest in automatic scoring (Nath *et al.*, 2023). Especially the problems that emerged in the evaluation processes with the Covid-19 pandemic (Şenel & Şenel, 2021) can be considered to have an important share in drawing the attention of researchers to this field.

Burrows *et al.* (2015) identified seven types of NLR for automatic scoring: (1) fill-in-the-blank, (2) short answer, (3) essay, (4) structured text, (5) maths, (6) source code, and (7) voice and speech techniques. The objective of this study is to implement an automated system for the evaluation of short-answer questions. Short answers consist of only a few words or sentences (Nath *et al.*, 2023). Burrows *et al.* (2015) distinguish short answers according to their length, focus, and clarity. In this answer type, the focus is on the meaning of the content. Short answers are objective and closed-ended in nature. The academic literature defines this domain as Automatic Short Answer Grading (ASAG). ASAG is a system that compares learner responses with one or more reference responses that are considered correct (Mohler & Mihalcea, 2009). Compared to manual systems, it can be said that these systems aim to provide justice more objectively (Badry *et al.*, 2023).

When the literature is analyzed, it can be stated that research in the field of ASAG has increased in recent years. Noticeably, English NLR datasets are used more frequently in the studies. However, developing pre-trained language models that can be used for many languages has facilitated research on different languages. Of particular interest are large language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-Trained Transformer), XLNet (Extra Long Network) and RoBERTa (Robustly optimised BERT approach), which also allow studies in multiple languages (Abdul-Mageed *et al.*, 2020; Zhang & Copus, 2023). This is because pre-trained models such as BERT can perform well on specialized tasks such as automatic scoring by leveraging large datasets to improve accuracy and efficiency (Chan *et al.*, 2024). The impact of LLMs is evident in the positive results observed in recent ASAG research on different languages (Mardini *et al.*, 2024; Sawatzki *et al.*, 2021; Sung *et al.*, 2019). These developments are promising for the dissemination of ASAG in different languages. This study aims to develop and evaluate two BERT-based ASAG models—an answer-based and a reference-based model—for the automatic scoring of Turkish short answers.

### 1.1. Related Work

The history of automatic scoring studies dates back to the 1960s (Page, 1967). With the developments in the field of NLP, interest in the ASAG field has recently increased. Different technological components have been utilized in the studies conducted in the historical process. Examples include word matching algorithms, similarity measures focusing on semantic distance, vector space representations, and machine learning algorithms. Burrows *et al.* (2015) classified the studies in the field of ASAG according to model approaches. The approaches are concept mapping, knowledge extraction, corpus-based, and machine learning. In another study, ASAG approaches are considered as similarity-oriented: (1) string-based, (2) semantic-based, (3) hybrid-based, and (4) machine and deep learning (Abdul-Salam *et al.*, 2022). It can be pointed out that attention is paid to historical development processes in approach classifications. In recent years, it is noteworthy that ASAG models have been mainly developed with a deep learning approach (Chaudhari & Patel, 2024). Research combining LSTM and derivative models with large language models (LLM), such as BERT, has achieved successful results (Gomaa *et al.*, 2023; Li *et al.*, 2022).

In line with the focus of the research, some end-use applications using the pre-trained BERT model were examined in the literature. We focused on BERT studies on different languages in the literature and general ASAG studies on Turkish. Zhu *et al.* (2022) developed a pre-trained BERT-based neural network model for the ASAG system. Their research included a semantic refinement layer consisting of Bi-LSTM (Bidirectional Long Short-Term Memory) and Capsule networks to improve the meaning of BERT outputs. The findings indicate that the developed model successfully obtained favorable outcomes compared to most other techniques and methods across the SemEval-2013 and Mohler datasets. In another study on SemEval-2013 tasks, BERT and XLNET pre-trained deep learning models were applied to the ASAG system

(Ghavidel *et al.*, 2020). In both models, improved results were achieved compared to previous studies. Sung *et al.* (2019) concentrated on advancing the existing BERT model by incorporating texts from related subject areas. They also created a second model through fine-tuning, which considered the student and reference answer pair for additional answer prediction. They designed a dataset of 3 sub-topics in the industrial field. The study's experimental results demonstrate that the BERT model, developed with text-based content during its training phase, outperformed the fine-tuned model. Amur *et al.* (2022) applied a BERT-based ASAG application with sQuAD 2.0 dataset to students in Roshan Tara school in Mehrabpur, Pakistan. The score predictions of the system were found to be highly successful.

Mardini *et al.* (2024) developed a dataset in Spanish targeting university students across ten subject areas, comprising 3772 answers scored from 0 (incorrect) to 5 (correct) and from 0 (incorrect) to 1 (correct). The study constructed the BERT model in six distinct approaches, and the answers were evaluated in both English and Spanish. In addition, the Skip-thought approach was also employed in the study. The study results indicated the potential contribution of the fine-tuned BERT model in developing reading comprehension skills in various languages. Nath *et al.* (2023) conducted experiments using the CREG (Meurers *et al.*, 2011) and CSSAG (Padó, 2016) datasets in German to develop an ASAG system. The BERT model yielded better results compared to the similarity-based approach utilizing bag-of-words. Notably, the CREG dataset provided more effective outcomes. Sawatzki *et al.* (2021) achieved highly successful outcomes compared to previous studies by employing a similar BERT approach for German and English datasets. The study used feature extraction architecture and fine-tuning with datasets from the Business Administration undergraduate program and the University of North Texas.

Nael *et al.* (2022) conducted one of the initial BERT studies on the Arabic language. They employed the Arabic adaptation of the Kaggle ASAP dataset to produce an ASAG system based on deep learning. Their study compared two novel approach models, BERT and ELECTRA (Efficiently Learning an Encoder That Accurately Classifies Token Replacements), with conventional deep learning models. The outcomes indicated that applying new approaches led to better results for short-answer scoring models. Schleifer *et al.* (2023) conducted a comparative study to ascertain superior performance. The AlephBERT PLM (Seker *et al.*, 2022) was employed to scrutinise Hebrew and the ASAG system was created with a dataset covering Biology topics. A performance comparison was, then, conducted between the AlephBERT-based and Convolutional Neural Network (CNN)-based systems. The study revealed that the AlephBERT-based model outperformed the CNN-based model.

Studies in the literature show that it is possible to develop ASAG systems with increased efficiency and reduced workload in various languages. Creating ASAG systems using traditional models typically demands in-depth feature engineering and considerable fine-tuning (Salim *et al.*, 2022). The chance to create more effective systems without lengthy NLP processes has prompted researchers to turn to pre-trained models like BERT (Chen *et al.*, 2023; Haller *et al.*, 2022). Analysis of studies indicates that BERT deep learning models have gained widespread adoption in the ASAG field recently. The data sets used in the systems developed with the BERT model and the results obtained are summarized in Table 1.

When examining the Turkish ASAG literature, it becomes apparent that only a limited number of studies have been conducted. To address this gap, Çınar *et al.* (2020) sought to develop a machine learning-based system to score Turkish short answers automatically. They collected a dataset containing answers from university-level physics students. SVM (Support Vector Machines), Gini, KNN (k-Nearest Neighbours), Bagging and Boosting techniques were used for model development. The highest model performance was obtained with AdaBoost.M1. This study drew attention as one of the pioneering studies in Turkish ASAG. The study conducted by Uysal and Doğan (2021) also consisted of short-answer items. The dataset comprised limited open-ended answers in the field of Turkish from the ABIDE program conducted by the Ministry



of National Education. Restricted open-ended items can be seen as equivalent to short-answer items. In their study, deep learning models were also used along with machine learning. The study evaluated the consistency between the automatic scoring mechanism created in the study and the scores provided by experts. The automatic scoring system development process employed five algorithms: SVM, LR (Logistic Regression), MNB (Multinomial Naive Bayes), LSTM, and BLSTM. Successful findings were recorded in this study, encouraging further investigation into automatic scoring studies on Turkish.

**Table 1.** Performance results of ASAG models developed with BERT.

Author	Dataset	Results
Sung <i>et al.</i> (2019)	SemEval-2013	Accuracy = .759, F1 = .758
Ghavidel <i>et al.</i> (2020)	SemEval-2013	Accuracy = .798, F1 = .797
Sawatzki <i>et al.</i> (2021)	German dataset	Quadratic weighted kappa (QWK) = .82, $r = .892$
Amur <i>et al.</i> (2022)	SQuad 2.0	QWK = .77, F1 = .96, Precision = .95
Nael <i>et al.</i> (2022)	ASAP-SAS	QWK = .77
Zhu <i>et al.</i> (2022)	SemEval-2013&Mohler	QWK = .82, $r = .892$
Schleifer <i>et al.</i> (2023)	Biology dataset	QWK $\geq .90$
Mardini <i>et al.</i> (2024)	Spanish dataset	RMSE = .59, $r = .78$

The available literature on Turkish ASAG studies indicates their limited number. Upon analysis of the existing studies, it is evident that traditional machine learning and deep learning models are prominent in the system development process. Distinct from these studies, our research aims to develop the Turkish ASAG system using BERT, a popular pre-trained deep learning model in the field. In this context, our investigation represents one of the pioneering endeavors in establishing a Turkish ASAG system utilizing BERT.

When the literature is examined, it is seen that answer-based and reference-based approaches stand out in the BERT model development process. In the answer-based approach, the training process is carried out by establishing a relationship between the Student Answer (SA) and the Expert Scores (ES) defined for all SAs. In the reference-based approach, the training process is carried out by establishing a similarity relationship between the SA and the Reference Answer (RA) (Nael *et al.*, 2022). Both approaches were considered in our study, and the automatic scoring of short Turkish answers was emphasized. This research constituted one of the pioneering studies in the field of automatic scoring of short answers in Turkish (ASAG) based on the BERT deep learning model, which has shown great success in recent years. The fact that BERT-based ASAG applications in Turkish have not yet been sufficiently covered in the literature increases the potential of this study to fill an important gap in the field and to serve as a basis for future research.

This study aims to develop and evaluate two BERT-based ASAG models—an answer-based and a reference-based model—for the automatic scoring of Turkish short answers. The research questions determined for the study in line with the research purpose are as follows;

- What is the scoring performance of BERT-based, answer-based and reference-based models in automatically evaluating Turkish short answers?
- How does the cognitive complexity of the assessment task (e.g., binary vs. multi-class, lower vs. higher-order thinking) influence the comparative performance of these models?

## 2. METHOD

Concurrently with ongoing research in this field, we incorporated the pre-existing BERT deep learning model in developing a Turkish ASAG system. The study utilized a comparative experimental research approach to evaluate two training methods for the BERT deep learning model to improve a Turkish ASAG system. Consequently, the results obtained from the



developed models were analyzed and interpreted to determine which method was optimal for development.

## 2.1. Dataset

In the study, the Physics data set was developed by Çınar *et al.* (2020). Two hundred forty-six (246) students studying in the Physics department at a state university responded to 4 physics-related questions. Inter-rater reliability was ensured by providing the dataset format for the answers evaluated by three experts (Çınar *et al.*, 2020). Inter-rater reliability values were calculated by the Pearson Correlation Coefficient. The correlations between the scores given by the raters to the short answer questions were .87 for question 1, .79 for question 2, .92 for question 3, .90 for question 4 and .87 for the average correlation. The average reliability correlation value of all questions was calculated by averaging the correlation values of the questions. The dataset contained student answers, reference answers, and expert scores. The reference answers were created as scoring keys. The question items and scoring keys in the data set were presented clearly in [Appendix 1](#). The study of Çınar *et al.* (2020) provided more detailed information about the question items.

Answers for Q1, Q3, and Q4 were objectively evaluated on a binary scale, while answers for Q2 received scores ranging from 0 to 4. The characteristics of the questions in the data set are presented in [Table 2](#), while [Table 3](#) shows the base data set properties.

**Table 2.** *Question characteristics.*

Question ID	Topic	Bloom's Taxonomy	Scoring Type
Q1	Electricity	Comprehension level	0-1
Q2	Conservation of energy	Comprehension level	0-1-2-3-4
Q3	Energy	Knowledge stage	0-1
Q4	Work	Knowledge stage	0-1

When the content of the dataset was analyzed, it was observed that it focused on core topics in physics, including electricity, energy, and work. The question structures, when classified according to Bloom's Taxonomy, aligned with the lower cognitive levels—specifically, knowledge and comprehension. However, despite all questions falling under these categories, they exhibited varying degrees of cognitive complexity in terms of the type of reasoning and elaboration required. For example, while Q1, Q3, and Q4 required relatively straightforward recall or identification of concepts (and were scored using binary scoring: 0 or 1), Q2 was designed to assess a more nuanced understanding and explanation of scientific reasoning, and thus used a multi-level scoring system (0–4). This task-based variability enabled us to examine how model performance was affected by differences in scoring structure and cognitive complexity, even within the same taxonomy level.

**Table 3.** *Base Physics dataset properties.*

Question	Answer Number	Distribution of Scores*	Scoring Type
Q1	254	89 / 165	0-1
Q2	147	73 / 6 / 8 / 24 / 36	0-1-2-3-4
Q3	254	31 / 223	0-1
Q4	254	155 / 99	0-1

\*The distribution of scores for Q1, Q3, and Q4 indicates the number of labels 0 and 1, and for Q2 indicates the number of labels 0, 1, 2, 3, and 4, respectively.

When the dataset was analyzed in detail in [Table 3](#), it was seen that the class distributions were significantly imbalanced. In the Distribution of Scores column, the scores given to the answers were presented in order. For Q1, there were 165 correct answers, while the number of incorrect answers was 89. In Q3, the number of incorrect answers was 31, while the number of correct

answers was 223. In Q2, it was observed that the number of collective responses needed to be higher, and the Distribution of the number of responses from the classes needed to be more balanced since multiple scoring type was used. Q4 was found to have a relatively more balanced scoring distribution than the others. For Q4, the number of incorrect answers was 155 while the number of correct answers was 99.

Using this dataset for the first time, Çınar *et al.* (2020) found a solution to the imbalanced dataset problem by using the Smote (Synthetic Minority Over-sampling) over-sampling technique. In this study, generative artificial intelligence (GenAI) technologies were used to balance the classification distributions in the dataset. The problem of class imbalance in the dataset was based on a sampling approach with equal representation of subgroups proposed by Zhang *et al.* (2024). As a result, the derived Physics dataset was ready for application in the study and the class imbalance problem had been solved. Table 4 illustrates the properties of the updated dataset.

**Table 4.** Updated Physics dataset properties.

Question	Answer Number	Distribution of Scores*	Scoring Type
Q1	265	100 / 165	0-1
Q2	266	66 / 50 / 50 / 50 / 50	0-1-2-3-4
Q3	274	100 / 174	0-1
Q4	254	150 / 128	0-1

\*The distribution of scores for Q1, Q3, and Q4 indicates the number of labels 0 and 1, and for Q2 indicates the number of labels 0, 1, 2, 3, and 4, respectively.

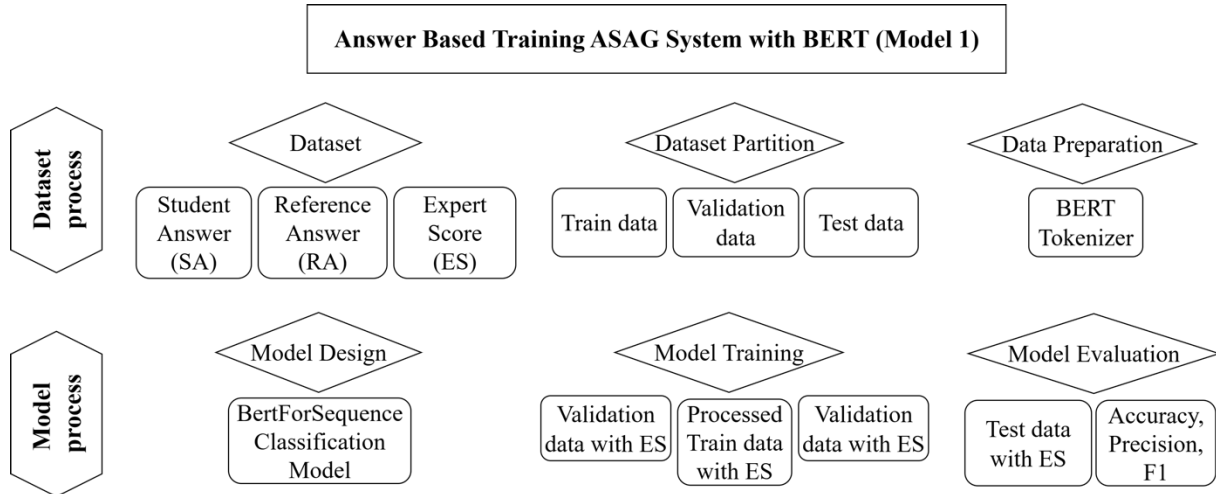
Empty answers in the Base Dataset were removed, and new answers that resembled student answers were derived, especially to balance the class distribution. The dataset had undergone minor modifications. The objectives of these actions were to enable the developed models to categorize the classes accurately and to equilibrate the class distribution in data classification partially. No data duplication via repetition was performed during data derivation. The purpose of these procedures was to ensure that the class distribution in the data set was balanced.

## 2.2. Model Design

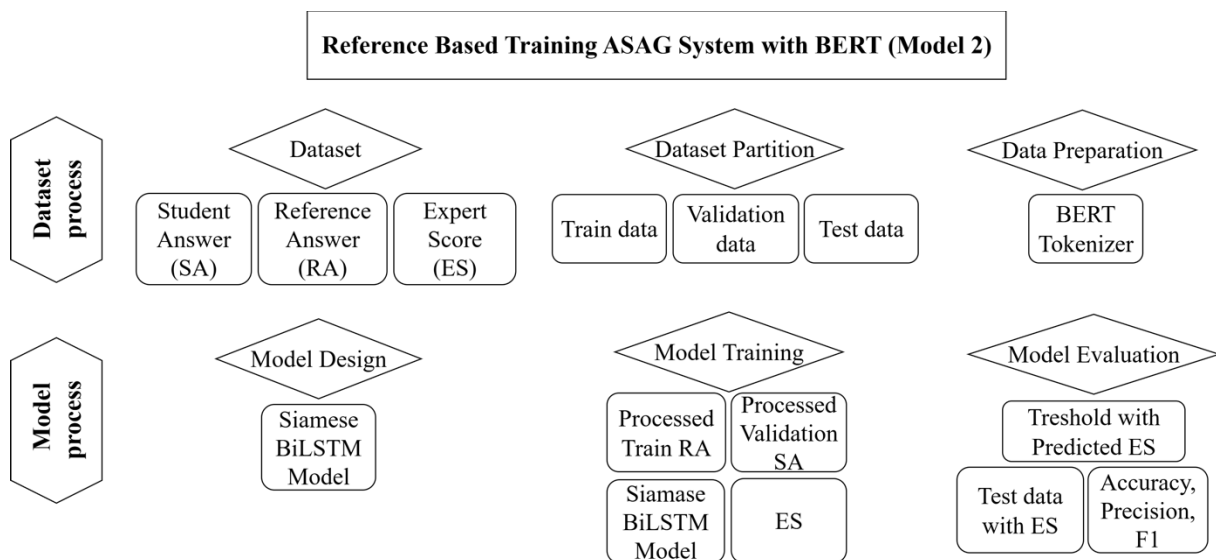
In the study, the “dbmdz/bert-base-turkish-uncased” model shared on the HuggingFace platform was preferred for automatic scoring of Turkish short answers. This model is a BERT model trained on large-scale text corpora, considering the specific features of Turkish language structure, and is considered a base model with high performance in Turkish natural language processing tasks. The “dbmdz/bert-base-turkish-uncased” model follows the original BERT architecture (Devlin *et al.*, 2019), including 12 transformer layers, 768-dimensional hidden layers, and 12 attention heads. The model contains approximately 110 million parameters and is pre-trained on 32GB of Turkish text.

The first method used student answers (SA) and expert scores (ES) for model training. The second method used the similarity between SA and reference answers (RA) to train the model. This study resulted in the development of two models suitable for both answer-based and reference-based training approaches. Figures 1 and 2 depict the general structure of the models designed for the automatic scoring of concise answers. The structures of both models are remarkably similar. They comprise the dataset and model processes indicated in Figures 1 and 2. Model 1, developed through answer-based training, has been created using expert scores. Model 2, developed through reference-based training, employs a similarity-based approach to system predictions. The dataset processing of both models followed similar stages. The dataset was initially selected and partitioned into training (70%), validation (15%), and test (15%) sets for Model 1. For Model 2, the same partitioning approach was applied with training (70%), validation (15%), and test (15%) subsets. BertTokenizer, a large language model created for the Turkish language, was implemented in the tokenization process for the models. Therefore,

the text data was transformed into a format appropriate for the BERT model. Model 1 used the Classification (CLS) embedding vector to derive features from all processed text inputs. CLS was particularly noteworthy in the realm of text classification (Devlin *et al.*, 2019; Sun *et al.*, 2019). This element contributed to the effectiveness of considerable bidirectional language models like BERT (Yang *et al.*, 2022).



**Figure 1.** Answer-based training ASAG system with BERT (Model 1).



**Figure 2.** Reference-based training ASAG system with BERT (Model 2).

The value information for the basic parameters we used in the model development process is presented in Table 5.

**Table 5.** Parameter values

Parameter	Value
Max length	128
Learning rate	2e−5
Batch size	32
Epochs	10
Early stopping patience	3
Dropout probability	.1

During the modeling process, the two approaches diverged in their training methods. Model 1 utilized the "BertForSequenceClassification" class to categorize text data for the Turkish BERT

model. Training directly aligned expert scores with student answers in the processed training set. In Model 2, the development process of the BERT model was executed utilizing Siamese BiLSTM. The model produced a similarity score between the answers of the students and those of the reference. The loss between the acquired similarity scores and expert scores was calculated to update the model. Adam optimizer and an epoch size of 10 were selected for optimization.

With large data sets, developing models with high-accuracy performance without overfitting problems was more practical. We aimed to overcome the disadvantage caused by the small size of our dataset with as simple a model setup as possible. We divided our dataset into training, testing, and validation at random intervals. Thus, we could directly monitor validation losses during training to avoid overfitting the models. The model training was terminated with the early stopping function when the performance stabilized. In our study, we also included the dropout technique to reduce the overfitting tendencies of the models. The dropout technique allowed model training to occur in multiple ways. Because some of the neurons were randomly deactivated in each training phase. The use of dropout strengthened the flexibility and generalizability of our models.

### 2.3. Data Analysis

The model development process and performance analysis of the developed models were carried out through Google Colab. The runtime type was “python 3”. “A100 GPU” was preferred for virtual GPU usage. Various evaluation metrics were applied to determine the effectiveness of ASAG models (Zesch *et al.*, 2023). In research studies, metrics such as accuracy, agreement, and correlation are commonly employed (Burrows *et al.*, 2015). Classification models commonly measure their performance using accuracy, precision, and F1 scores (Chaudhari & Patel, 2024). Accuracy is the ratio of correct model predictions to total answers, while precision is concerned with the number of true and false positives. The F1 score is the harmonic mean of precision and recall values. It is generally preferable to see a better model performance in cases with unbalanced class distributions (Riyanto *et al.*, 2023). These metrics were used to evaluate the performance of the developed models.

In addition to these individual performance metrics, a direct statistical comparison between Model 1 and Model 2 was conducted to identify significant differences in their predictions. For this purpose, a dual-analysis approach was adopted. First, McNemar's test was employed. This non-parametric test is specifically designed for paired nominal data and is used to determine if the two models have differing error rates. It evaluates whether one model is significantly more likely to be correct when the other is incorrect. Second, Cohen's Kappa ( $\kappa$ ) was calculated as a measure of inter-rater agreement between each model and the expert scores. Unlike simple accuracy, Cohen's Kappa accounts for the possibility of agreement occurring by chance, providing a more robust measure of performance, especially in multi-class or imbalanced datasets. All statistical analyses were performed using the Scikit-learn library in Python.

The data analysis also included a qualitative error analysis to complement the quantitative findings. This analysis focused on instances where the models' predictions diverged from the expert scores, particularly on the more complex, multi-class task (Q2). The objective of this analysis was to identify systematic error patterns, understand the types of student answers that would pose challenges for the models, and gain deeper insights into the models' decision-making processes. The process involved a manual review of misclassified responses to categorize the nature of the errors. The implementation of all statistical and qualitative analyses relied on the Scikit-learn and Pandas libraries in Python.

## 3. RESULTS

Model 1 was the ASAG model, developed through an answer-based learning approach. The performance of each request within the dataset was extracted and measured using accuracy,

precision, and F1 score. The corresponding results for Model 1's performance are presented in Table 6.

**Table 6.** Answer-based (Model 1) ASAG performance.

Question	Accuracy	Precision*	F1 Score**
Q1	.82	.82 / .82	.82 / .82
Q2	.90	.91 / .91	.90 / .90
Q3	.94	.96 / .92	.93 / .94
Q4	.86	.86 / .86	.86 / .86

\*Precision values represent both macro and weighted values.

\*\*F1 Score values also represent both macro and weighted values.

Upon analyzing the performance results of Model 1, it can be concluded that the model successfully dealt with the physics dataset obtained. Accuracy scores ranging from .82-.94 were achieved for all questions by the Accuracy metric. The Precision results were also promising, with macro scores of .82-.96 and weighted scores of .82-.92. While macro average values for F1 scores were .82-.93, weighted values were .82-.94. The main difference between macro and weighted average is that the macro average gives equal importance to each class, while the weighted average takes into account class imbalance. The macro average calculates and averages the metric for each class independently. Weighted average calculates the metric for each class but weights it according to the number of samples in that class. In the context of this study, the macro precision/F1 score gave equal importance to each scoring category, while the weighted precision/F1 score gave more importance to scoring categories with more student responses.

Model 2 represents the ASAG model, developed using a reference-based learning approach. The prediction results were assessed utilizing accuracy, precision, and F1 scores, much like in Model 1. The performance data of Model 2 are presented in Table 7.

**Table 7.** Reference-based (Model 2) ASAG performance.

Question	Accuracy	Precision*	F1 Score**
Q1	.85	.90 / .88	.83 / .84
Q2	.62	.64 / .62	.61 / .57
Q3	.95	.94 / .95	.94 / .95
Q4	.86	.86 / .86	.86 / .86

\*Precision values represent both macro and weighted values.

\*\*F1 Score values also represent both macro and weighted values.

The results of Model 2 also yielded a satisfactory outcome for this dataset. Accuracy scores ranged from .62 to .95. Precision macro values ranged from .64 to .94, while weighted average values ranged from .62 to .95. F1 scores for macro were between .61 and .94, while weighted average values ranged from .57 to .95. As shown in Table 7, the accuracy and precision values were in harmony. Thus, the accuracy performance of Model 2 can be generalized to the new data and is not subject to overfitting.

The performance metrics for Model 2, the reference-based ASAG model, are presented in Table 7. The model's performance varied significantly across the different tasks. Accuracy scores ranged from a low of .62 to a high of .95. A similar spread was observed in the F1 scores, with weighted F1 scores spanning from .57 to .95. Notably, the model's lowest performance metrics were consistently recorded on the multi-class scoring task (Q2), with an accuracy of .62 and a weighted F1 score of .57.

A preliminary review of Tables 6 and 7 indicates that both models perform effectively on the dataset, but a more direct and statistically robust comparison is necessary to discern significant



performance differences. To achieve this, a formal statistical comparison was conducted using a dual-analysis approach. First, McNemar's test was employed to identify statistically significant differences in the error rates of the two models. Second, Cohen's Kappa ( $\kappa$ ) was calculated to provide a more nuanced measure of agreement between each model's predictions and the expert scores, which is particularly suitable for tasks with multiple or ordinal scoring categories. The comprehensive results of this comparative analysis are presented in Table 8.

**Table 8.** Comprehensive comparison of model performance.

Question	Task Type	McNemar $p$ -value	Superior Model	Model 1 Kappa ( $\kappa_1$ )	Model 2 Kappa ( $\kappa_2$ )
Q1	Binary	< .001	Model 2	.575	.681
Q2	Multiple	< .001	Model 1	.843	.532
Q3	Binary	> .999	n.s	.952	.952
Q4	Binary	> .999	n.s*	.631	.678

\*n.s. = not significant. While the Kappa scores for Q4 show a numerical difference, the McNemar test indicates that this difference is not statistically significant.

For the multi-class scoring task (Q2), a statistically significant difference in performance was identified (McNemar's test,  $p < .001$ ). Model 1 demonstrated a substantially higher agreement with expert scores ( $\kappa_1 = .843$ ) compared to Model 2 ( $\kappa_2 = .532$ ). A significant difference was also found for question Q1 (McNemar's test,  $p < .001$ ); however, in this case, Model 2 achieved a higher agreement score ( $\kappa_2 = .681$ ) than Model 1 ( $\kappa_1 = .575$ ). Finally, for the binary classification tasks Q3 and Q4, no statistically significant difference was detected between the models ( $p > .999$ ). Their Cohen's Kappa scores were also highly comparable for these questions ( $\kappa_1 = .952$  vs.  $\kappa_2 = .952$  for Q3, and  $\kappa_1 = .631$  vs.  $\kappa_2 = .678$  for Q4).

To further investigate the quantitative performance of Model 2, particularly its low agreement score ( $\kappa = .532$ ) on the multi-class task (Q2), a qualitative error analysis was conducted on its predictions. The analysis revealed several systematic error patterns. The most prominent error pattern was the systematic misclassification of answers with a true score of '2', which were nearly all incorrectly assigned a score of '0'. These misclassified responses were typically characterized by their conciseness. While they correctly stated the core scientific principle, it was observed that their presentation was very direct. Examples of such answers include:

- *Hızları değişmez hız kütleyle bağlı değildir [The speeds do not change; speed is not dependent on mass].*
- *Kütlenin düşme hızı bağlamında bir etkisi yoktur, aynı hızda düşerler [Mass has no effect in the context of falling speed; they fall at the same speed].*
- *Aynıdır serbest düşme mantığına göre kütlenin bir önemi yoktur [It is the same; according to the logic of free fall, mass is not important].*

A common feature of these answers is their lack of detailed explanations involving concepts such as potential or kinetic energy, which are present in higher-scoring responses and the reference answer.

A second observed error pattern is the model's inconsistent differentiation between scores of '3' and '4'. The model occasionally assigns a different score to answers that are textually and conceptually very similar. Finally, a notable outlier is an instance where the model assigns a score of '0' to a detailed, conceptually rich answer with a true score of '3'. This particular answer introduces a related concept ("air resistance") not present in the reference answer.

#### 4. DISCUSSION and CONCLUSION

This study developed and compared two BERT-based models for the automatic scoring of Turkish short answers: an answer-based (Model 1) and a reference-based (Model 2) model. Our results revealed that the choice of the optimal scoring model was not absolute but was contingent on the cognitive complexity of the assessment task.

#### 4.1. The Impact of Task Complexity on Model Performance

The central results of this research are the clear interaction between model architecture and task complexity, which manifested across a spectrum of assessment types. For the most cognitively demanding, multi-class task (Q2), the answer-based model (Model 1) was unequivocally superior, with a significantly higher agreement score than its counterpart ( $\kappa_1 = .843$  vs.  $\kappa_2 = .532$ ,  $p < .001$ ). This suggests that when answers require nuanced understanding and can be expressed in many valid ways, a model trained on a diverse set of student responses is better equipped to learn the complex patterns of partial and full credit. Conversely, for the medium-complexity task of identifying a specific, well-defined misconception (Q1), the reference-based model (Model 2) demonstrated a statistically significant advantage ( $\kappa_2 = .681$  vs.  $\kappa_1 = .575$ ,  $p < .001$ ). Similarly, Sayeed and Gupta (2022) emphasized that reference-based approaches demonstrated superior performance, particularly for medium and lower complexity tasks, achieving significant improvements in ASAG systems when utilizing Siamese-based transformers that model the evaluation as sentence similarity between reference and student answer pairs. Finally, for the low-complexity definitional tasks (Q3 & Q4), both models performed at a high level with no statistically significant difference, suggesting the architectural choice was less critical for simple knowledge recall. These results further validated the observations by Zhu *et al.* (2022), who emphasized that fine-tuned BERT models could achieve successful results even with small corpora, demonstrating the robustness of transformer-based architectures across varying dataset sizes and task complexities. Similarly, Salim *et al.* (2022) emphasized that pre-trained BERT models, through their transfer learning capabilities and effective fine-tuning processes, could achieve high performance even with limited data.

#### 4.2. Understanding Model Limitations: A Qualitative Perspective

The reference-based model (Model 2), while effective in binary classification tasks, exhibited a significant limitation in the multi-class scoring task, particularly in its complete inability to identify answers corresponding to the intermediate 'Score 2' category. Qualitative error analysis suggests this issue stems from the model's fundamental design, which equates semantic similarity to a single, high-quality reference answer with scoring accuracy. This approach systematically penalized answers that were conceptually correct but concise, as they lacked the detailed phrasing and specific keywords present in the comprehensive reference answer. Intermediate scores may have become an ambiguous 'middle ground' that the model struggled to learn. In this situation, the model's struggle to differentiate between a student's grasp of the core concept and their ability to articulate it in a textually similar manner may stem from a key vulnerability inherent in single-reference-based approaches for nuanced, multi-level assessment scenarios. A potential solution to this problem is to integrate multiple reference answers representing diverse yet pedagogically valid expressions. This may help the model capture a broader semantic space and reduce the penalization of concise but correct answers. Similarly, Akila-Devi *et al.* (2023) emphasized that reference-based approaches significantly enhanced performance in ASAG systems, particularly when reference answers were strengthened through diverse content acquisition from multiple sources, including expert responses and community question-answering platforms. Indeed, implementing a strategy such as clustering student responses thematically and using centroid-based representations as auxiliary references could support better outcomes, improving the model's sensitivity to variation in student phrasing, especially in mid-range scoring categories.

The challenges we observed with our reference-based model's inability to handle intermediate scoring categories could be attributed to dataset imbalance or insufficient data issues, particularly affecting multi-class classification tasks that required nuanced scoring distinctions. Similarly, Mardini *et al.* (2024) emphasized in the literature that reference-based approaches

showed low performance in predicting extreme values due to imbalanced data distribution combined with insufficient data in multi-class classification tasks. In addition, employing semi-supervised learning techniques—such as pseudo-labeling or confidence-based refinement—could help the model learn more effectively from unlabeled or uncertain responses. This approach allowed the model to generate provisional labels for ambiguous student answers and used them in further training cycles. Specifically, it could help the model handle mid-range scores more accurately, which were often challenging not only due to underrepresentation but also because such answers tended to be concise or expressed in diverse, non-standard ways that deviated from the reference answer's phrasing. By incorporating these borderline or confidently predicted samples, the model can better be generalized across varying answer styles and levels of elaboration. Similarly, Xie *et al.* (2023) emphasized that pseudo-labeling enhanced class prediction accuracy, as demonstrated by their CAP (Class-Aware Pseudo-Labeling) method.

### 4.3. Contextualizing Performance Within the Field

The performance metrics achieved in our study are highly competitive, aligning closely with results from BERT-based ASAG systems in other languages such as German (Sawatzki *et al.*, 2021), Spanish (Mardini *et al.*, 2024) and Arabic (Nael *et al.*, 2022). This demonstrates that the effectiveness of transformer-based architectures for automatic scoring is not language-specific and that Turkish, as a morphologically rich language, can similarly benefit from these advanced models.

The advantages of cross-linguistic consistency, efficiency and generalizability became even clearer when comparing our results to the work of Çınar *et al.* (2020), who used the same dataset. M1 model achieved a marginally higher peak F1 score; our BERT models reached a comparable level of performance without the need for extensive, manual feature engineering. This distinction is critical, as traditional machine-learning approaches are known to be complex and require significant manual intervention (Burrows *et al.*, 2015; Zehner *et al.*, 2016). Therefore, the key advantage of our approach lies in its efficiency. By leveraging pre-trained models, we demonstrate a more direct and scalable pathway for developing high-performing ASAG systems. This result resonates with the broader trend in the field away from feature-dependent models (Abdul-Salam *et al.*, 2022) and can provide strong motivation for the continued development of BERT-based models for Turkish.

### 4.4. Limitations

It is important to acknowledge several limitations of this study, most of which pertain to the dataset and its scope. These limitations may influence the generalizability and interpretability of the findings, particularly in relation to model behavior across different tasks and contexts.

- **Domain Limitation:** The dataset was limited to a single subject domain—Physics—which may restrict the applicability of the results to other disciplines with different linguistic or conceptual characteristics.
- **Cognitive Level Limitation:** All questions in the dataset targeted only the knowledge and comprehension levels of Bloom's Taxonomy. Therefore, the findings may not extend to higher-order cognitive tasks such as application, analysis, or evaluation, where students' language use and reasoning strategies may differ substantially.
- **Data Size and Diversity:** The study involved a relatively small dataset, including only four questions and responses from 246 students. This modest scale may not capture the full variability present in real-world educational settings. In particular, multi-class classification tasks may require more data to accurately model nuanced score boundaries, as noted in recent literature (Mardini *et al.*, 2024).
- **Language-Specific Bias:** This study's exclusive focus was on Turkish—a morphologically rich and agglutinative language—may limit the generalizability of our findings to languages with different linguistic structures. Nonetheless, the strong performance of our models suggests that transformer-based architectures like BERT could effectively handle such

complexity. While this remains a limitation, it also positions Turkish as a promising test case that may encourage further research across diverse language families.

Given these limitations, caution should be exercised when interpreting the results beyond the context of this specific dataset. As the reliability and validity of models are best understood across different datasets (Zhu *et al.*, 2022), future work should aim to test the durability and sustainability of these models on a wider range of subjects and different cognitive tasks.

#### 4.5. Implications

In summary, this research demonstrates the effectiveness of BERT-based models for the automatic scoring of short answers in Turkish. An important further contribution of the study is the result that the optimal model architecture is not universal but is contingent on the cognitive complexity of the assessment task. Specifically, an answer-based approach excels for complex, multi-faceted questions, while a reference-based model is more reliable for identifying specific, well-defined concepts.

Furthermore, this research establishes that these pre-trained models can achieve performance levels comparable to traditional machine learning techniques while significantly reducing the need for laborious feature engineering. This can provide a powerful and potentially more efficient pathway for developing ASAG systems for Turkish and make a significant contribution to a field where such applications are limited. While the promising performance indicates significant potential for practical applications, the generalizability of these findings requires further research using larger and more diverse datasets.

#### 4.6. Recommendations and Future Work

##### 4.6.1. Model-Specific Recommendations

In light of the results obtained from this study—particularly the limitations observed in the reference-based model during multi-class scoring—several potential improvements can be considered to enhance model performance in similar contexts:

- Incorporating multiple reference answers may help the model better recognize semantically correct but textually diverse student responses, especially for intermediate score categories. By capturing a broader range of acceptable expressions, this approach could reduce the likelihood of penalizing valid but concise answers.
- Thematic clustering of student responses and the use of centroid-based representations as auxiliary references might contribute to improving the model's robustness against lexical and structural variation. This data-driven approach could complement manually created references by reflecting how students naturally express their understanding.
- Semi-supervised learning methods, such as pseudo-labeling or confidence-based refinement, may offer a way to utilize ambiguous or unlabeled responses more effectively. These techniques could support the model in learning from borderline or underrepresented cases and improve its ability to generalize in nuanced scoring tasks.

These strategies do not represent definitive solutions, but they may provide useful directions for addressing the challenges encountered with reference-based ASAG models in future work.

##### 4.6.2. Broader Research Directions

Beyond the technical aspects of the models, several broader areas of inquiry may also support the advancement of ASAG systems for Turkish and similar languages:

- New insights can be brought to the field by implementing the ASAG models in learning environments and taking the opinions of teachers and students regarding performance status.
- Different LLM models developed for Turkish can be compared with the performance of ASAG, and even more reliable performance outputs can be obtained by considering these models together.

- In the future, using automatic formative feedback systems with ASAG applications will strengthen the communication between students and educators in the follow-up of learning situations.
- The data set used in this study corresponds to only two thinking processes in Bloom's Taxonomy: knowledge and comprehension. In future studies, ASAG performances can be investigated on data sets for higher cognitive levels in Bloom's Taxonomy.
- Additional research may also examine ways to enhance the explainability of automatic scoring decisions, thereby increasing transparency and trust among educators and learners.

### Acknowledgments

We thank Çınar *et al.* (2020) for sharing the physics dataset as an open data source.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

### Contribution of Authors

**Abdulkadir Kara:** Investigation, Resources, Methodology, Software, Formal analysis, and Writing-original draft. **Zeynep Avinç Kara:** Investigation, Resources, Validation, and Formal analysis. **Serkan Yıldırım:** Methodology, Supervision, Validation, and Writing-original draft.

### Orcid

Abdulkadir Kara  <https://orcid.org/0000-0003-3255-1408>

Zeynep Avinç Kara  <https://orcid.org/0000-0002-8309-3876>

Serkan Yıldırım  <https://orcid.org/0000-0002-8277-5963>

### REFERENCES

- Abdul-Mageed, M., Elmadany, A., & Nagoudi, E.M.B. (2020). *ARBERT & MARBERT: Deep bidirectional transformers for Arabic*. arXiv. <https://doi.org/10.48550/arXiv.2101.01785>
- Abdul-Salam, M., El-Fatah, M.A., & Hassan, N.F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. *PloS ONE*, 17(8), Article e0272269. <https://doi.org/10.1371/journal.pone.0272269>
- Akila Devi, T.R., Javubar Sathick, K., Abdul Azeez Khan, A., & Arun Raj, L. (2023). Novel framework for improving the correctness of reference answers to enhance results of ASAG systems. *SN Computer Science*, 4(4), Article 415. <https://doi.org/10.1007/s42979-023-01682-8>
- Amur, Z.H., Hooi, Y.K., & Soomro, G.M. (2022). Automatic short answer grading (ASAG) using attention-based deep learning MODEL. In *2022 International Conference on Digital Transformation and Intelligence* (pp. 1-7). Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICDI57181.2022.10007187>
- Badry, R.M., Ali, M., Rslan, E., & Kaseb, M.R. (2023). Automatic arabic grading system for short answer questions. *IEEE Access*, 11, 39457-39465. <https://doi.org/10.1109/ACCESS.2023.3267407>
- Benli, I., & İsmailova, R. (2018). Use of open-ended questions in measurement and evaluation methods in distance education. *International Technology and Education Journal*, 2(1), 1-8.
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25, 60-117. <https://doi.org/10.1007/s40593-014-0026-8>
- Chan, S., Sathyamurthy, M., Inoue, C., Bax, M., Jones, J., & Oyekan, J. (2024). Integrating metadiscourse analysis with transformer-based models for enhancing construct representation and discourse competence assessment in l2 writing: A systemic



- multidisciplinary approach. *Journal of Measurement and Evaluation in Education and Psychology*, 15(Special Issue), 318-347. <https://doi.org/10.21031/epod.1531269>
- Chaudhari, R., & Patel, M. (2024). Deep learning in automatic short answer grading: A comprehensive review. *ITM Web of Conferences*, 65, Article 03003. <https://doi.org/10.1051/itmconf/20246503003>
- Chen, X., Zhou, Z., & Prado, M. (2025). ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams. *International Journal of Assessment Tools in Education*, 12(1), 62-77. <https://doi.org/10.21449/ijate.1496193>
- Chen, Y., Luo, J., Zhu, X., Wu, H., & Yuan, S. (2023). A cross-lingual hybrid neural network with interaction enhancement for grading short-answer texts. *IEEE Access*, 11, 37508-37514. <https://doi.org/10.1109/ACCESS.2023.3260840>
- Çınar, A., İnce, E., Gezer, M., & Yılmaz, O. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies*, 25(5), 3821-3844. <https://doi.org/10.1007/s10639-020-10128-0>
- Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- Dönmez, M. (2024). AI-based feedback tools in education: a comprehensive bibliometric analysis study. *International Journal of Assessment Tools in Education*, 11(4), 622-646. <https://doi.org/10.21449/ijate.1467476>
- Filighera, A., Ochs, S., Steuer, T., & Tregel, T. (2023). Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. *International Journal of Artificial Intelligence in Education*, 34, 616-646. <https://doi.org/10.1007/s40593-023-00361-2>
- Garg, J., Papreja, J., Apurva, K., & Jain, G. (2022). Domain-specific hybrid BERT based system for automatic short answer grading. In *Proceedings of 2<sup>nd</sup> International Conference on Intelligent Technologies* (pp. 1-6). The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/CONIT55038.2022.9847754>
- Ghavidel, H.A., Zouaq, A., & Desmarais, M.C. (2020). Using BERT and XLNET for the automatic short answer grading task. In H.C. Lane, S. Zvacek, & J. Uhomobhi (Eds.), *Proceedings of the 12th International Conference on Computer Supported Education - (Volume 1)* (pp. 58-67). SciTePress. <https://doi.org/10.5220/0009422400580067>
- Gomaa, W.H., Nagib, A.E., Saeed, M.M., Algarni, A., & Nabil, E. (2023). Empowering short answer grading: integrating transformer-based embeddings and BI-LSTM network. *Big Data and Cognitive Computing*, 7(3), Article 122. <https://doi.org/10.3390/bdcc7030122>
- Haller, S., Aldea, A., Seifert, C., & Strisciuglio, N. (2022). *Survey on automatic short answer grading with deep learning: From word embeddings to transformers*. arXiv. <https://doi.org/10.48550/arXiv.2204.03503>
- Hasanah, U., Permasari, A.E., Kusumawardani, S.S., & Pribadi, F.S. (2016, August). A review of an information extraction technique approach for automatic short answer grading. In *Proceedings of 2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering* (pp. 192-196). Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/ICITISEE.2016.7803072>
- Jadidinejad, A.H., & Mahmoudi, F. (2014). Unsupervised short answer grading using spreading activation over an associative network of concepts / la notation sans surveillance des réponses courtes en utilisant la diffusion d'activation dans un réseau associatif de concepts. *Canadian Journal of Information and Library Science*, 38(4), 287-303.
- Katsaris, I., & Vidakis, N. (2021). Adaptive e-learning systems through learning styles: a review of the literature. *Advances in Mobile Learning Educational Research*, 1(2), 124-145. <https://doi.org/10.25082/AMLER.2021.02.007>

- Kurbanoğlu, N.I., & Olcaytürk, M. (2023). Investigation of the exam question types attitude scale for secondary school students: development, validity, and reliability. *Sakarya University Journal of Education*, 13(2), 191-206. <https://doi.org/10.19126/suje.1187470>
- Leacock, C., & Chodorow, M. (2003). C-rater: Automatic scoring of short-answer questions. *Computers and the Humanities*, 37, 389-405. <https://doi.org/10.1023/A:1025779619903>
- Li, X., Li, X., Chen, S., Ma, S., & Xie, F. (2022). Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. *Connection Science*, 34(1), 1638-1653. <https://doi.org/10.1080/09540091.2022.2078279>
- Lubis, F.F., Putri, A., Waskita, D., Sulistyaningtyas, T., Arman, A.A., & Rosmansyah, Y. (2021). Automatic short-answer grading using semantic similarity based on word embedding. *International Journal of Technology*, 12(3), 571-581. <https://doi.org/10.14716/ijtech.v12i3.4651>
- Mardini, G.I.D., Quintero, M.C.G., Vilorio, N.C.A., Percybrooks, B.W.S., Robles, N.H.S., & Villalba, R.K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Education and Information Technologies*, 29(4), 4565-4590. <https://doi.org/10.1007/s10639-023-11890-7>
- Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In S. Padó, & S. Thater (Eds.), *Proceedings of the TextInfer 2011 workshop on textual entailment* (pp. 1–9). Association for Computational Linguistics. <https://aclanthology.org/W11-2401/>
- Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th Conference of the European chapter of the association for computational linguistics* (pp. 567-575). Association for Computational Linguistics. <https://aclanthology.org/E09-1065.pdf>
- Nael, O., ElManyalawy, Y., & Sharaf, N. (2022). AraScore: a deep learning-based system for Arabic short answer scoring. *Array*, 13, Article 100109. <https://doi.org/10.1016/j.array.2021.100109>
- Nath, S., Parsaeifard, B., & Werlen, E. (2023, August 22-26). *Automatic short answer grading using BERT on German datasets* [Paper presentation]. 20<sup>th</sup> Biennial EARLI Conference, Thessaloniki, Greece.
- Noyes, K., McKay, R.L., Neumann, M., Haudek, K.C., & Cooper, M.M. (2020). Developing computer resources to automate analysis of students' explanations of London dispersion forces. *Journal of Chemical Education*, 97(11), 3923-3936. <https://doi.org/10.1021/acs.jchemed.0c00445>
- Padó, U. (2016). Get semantic with me! the usefulness of different feature types for short-answer grading. In Y. Matsumoto, & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical Papers* (pp. 2186-2195). The COLING 2016 Organizing Committee. <https://aclanthology.org/C16-1206/>
- Page, E.B. (1967). Grading essays by computer: Progress report. *Proceedings of the Invitational Conference on Testing Problems*, 87-100.
- Ramineni, C., & Williamson, D.M. (2013). Automatic essay scoring: psychometric guidelines and practices. *Assessing Writing*, 18(1), 25-39. <https://doi.org/10.1016/j.asw.2012.10.004>
- Riyanto, S., Imas, S.S., Djatna, T., & Atikah, T.D. (2023). Comparative analysis using various performance metrics in imbalanced data for multi-class text classification. *International Journal of Advanced Computer Science and Applications*, 14(6). <https://doi.org/10.14569/IJACSA.2023.01406116>
- Salim, H.R., De, C., Pratamaputra, N.D., & Suhartono, D. (2022). Indonesian automatic short answer grading system. *Bulletin of Electrical Engineering and Informatics*, 11(3), 1586-1603. <https://doi.org/10.11591/eei.v11i3.3531>
- Saunders, D.R., Bex, P.J., Rose, D.J., & Woods, R.L. (2014). Measuring information acquisition from sensory input using automatic scoring of natural-language

- descriptions. *PLoS ONE*, 9(4), Article e93251. <https://doi.org/10.1371/journal.pone.0093251>
- Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2021). Deep learning techniques for automatic short answer grading: predicting scores for English and German answers. In E. Cheng, R.B. Koul, T. Wang, & X. Yu (Eds.), *Proceedings of 2021 2<sup>nd</sup> international conference on artificial intelligence in education technology* (pp. 65-75). Springer. [https://doi.org/10.1007/978-981-16-7527-0\\_5](https://doi.org/10.1007/978-981-16-7527-0_5)
- Sayeed, M.A., & Gupta, D. (2022). Automate descriptive answer grading using reference based models. In *Proceedings of 2022 OITS international conference on information technology* (pp. 262-267). The Institute of Electrical and Electronics Engineers. <https://doi.org/10.1109/OCIT56763.2022.00057>
- Schleifer, A.G., Klebanov, B.B., Ariely, M., & Alexandron, G. (2023). Transformer-based Hebrew NLP models for short answer scoring in biology. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications* (pp. 550-555). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bea-1.46>
- Seker, A., Bandel, E., Bareket, D., Brusilovsky, I., Greenfeld, R., & Tsarfaty, R. (2022). AlephBERT: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 46-56). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.4>
- Siddiqi, R., Harrison, C.J., & Siddiqi, R. (2010). Improving teaching and learning through automatic short-answer marking. *IEEE Transactions on Learning Technologies*, 3(3), 237-249. <https://doi.org/10.1109/TLT.2010.4>
- Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019, November). Pre-training BERT on domain resources for short answer grading. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 6071-6075). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1628>
- Şenel, S., & Şenel, H.C. (2021). Remote assessment in higher education during COVID-19 pandemic. *International Journal of Assessment Tools in Education*, 8(2), 181-199. <https://doi.org/10.21449/ijate.820140>
- Tulu, C.N., Özkaya, O., & Orhan, U. (2021). Automatic short answer grading with semspace sense vectors and malstm. *IEEE Access*, 9, 19270-19280. <https://doi.org/10.1109/ACCESS.2021.3054346>
- Uto, M., & Uchida, Y. (2020). Automatic short-answer grading using deep neural networks and item response theory. In I.I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millan (Eds.), *Proceedings of the 21th International Conference on Artificial Intelligence in Education* (pp. 334-339). Springer International Publishing. [https://doi.org/10.1007/978-3-030-52240-7\\_61](https://doi.org/10.1007/978-3-030-52240-7_61)
- Uyar, A.C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: a focus on ChatGPT. *International Journal of Assessment Tools in Education*, 12(1), 20-32. <https://doi.org/10.21449/ijate.1517994>
- Uysal, I., & Doğan, N. (2021). How reliable is it to automatically score open-ended items? An application in the Turkish language. *Journal of Measurement and Evaluation in Education and Psychology*, 12(1), 28-53. [https://doi.org/10.1007/978-3-030-52240-7\\_61](https://doi.org/10.1007/978-3-030-52240-7_61)
- Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automatic essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers & Education*, 123, 212-224. <https://doi.org/10.1016/j.compedu.2018.05.010>
- Xie, M.K., Xiao, J., Liu, H.Z., Niu, G., Sugiyama, M., & Huang, S.J. (2023). Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *Advances in Neural*

- Information Processing Systems*, 36, 25731-25747. <https://doi.org/10.48550/arXiv.2305.02795>
- Yang, X., Huang, J.Y., Zhou, W., & Chen, M. (2022). *Parameter-efficient tuning with special token adaptation*. arXiv. <https://doi.org/10.48550/arXiv.2210.04382>
- Yıldırım, O., & Demir, S.B. (2022). Inside the black box: Do teachers practice assessment as learning?. *International Journal of Assessment Tools in Education*, 9(Special Issue), 46-71. <https://doi.org/10.21449/ijate.1132923>
- Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, 76(2), 280-303. <https://doi.org/10.1177/0013164415590022>
- Zesch, T., Horbach, A., & Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice*, 42(1), 44-58. <https://doi.org/10.1111/emip.12544>
- Zhang, L., & Copus, B. (2023). A Study of Compressed Language Models in Social Media Domain. *The International FLAIRS Conference Proceedings*, 36(1). <https://doi.org/10.32473/flairs.36.133056>
- Zhang, M., Johnson, M., & Ruan, C. (2024). Investigating sampling impacts on an LLM-based AI scoring approach: Prediction accuracy and fairness. *Journal of Measurement and Evaluation in Education and Psychology*, 15(Special Issue), 348-360. <https://doi.org/10.21031/epod.1561580>
- Zhu, X., Wu, H., & Zhang, L. (2022). Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies*, 15(3), 364-375. <https://doi.org/10.1109/tlt.2022.3175537>
- Zimmerman, W.A., Kang, H.B., Kim, K., Gao, M., Johnson, G., Clariana, R., & Zhang, F. (2018). Computer-automatic approach for scoring short essays in an introductory statistics course. *Journal of Statistics Education*, 26(1), 40-47. <https://doi.org/10.1080/10691898.2018.1443047>



## APPENDIX


### Appendix 1. Physics dataset questions and scoring keys (Çınar et al., 2020).

Question ID	Question Text	Scoring Key
Q1	<p>Sinan is preparing the report of the experiment they did in the physics laboratory class that day. The question in the test report is:</p> <p>“If there is a single bulb in an electrical circuit, in which case two bulbs of the same power are connected in series, compare the brightness of the bulbs and explain the reason for your answer.”</p> <p>Sinan rethinks their experiments in the laboratory and answers the question as follows:</p> <p>“While there is only one light bulb in an electrical circuit, the light bulb is quite bright. When a second bulb is added serially to this bulb, both bulbs light up equally and less brightly. The reason for this is that the single bulb uses all the current flowing through the circuit. When there are two bulbs in the circuit, the current is shared by the bulbs and therefore they are equal but less bright.”</p> <p>Sinan thinks he understands very well what is happening in the experiment, but he makes a big mistake. Can you find this error?</p>	<p>0 point: If student did not answer or gave an incorrect answer.</p> <p>1 point: Sinan thinks that the current passing through both bulbs and the bulbs share this current. However, the same current passes first through the first bulb and then through the other. The two bulbs are dimmer than a single bulb.</p> <p>Because the increased resistance in the circuit reduces the current.</p> <p>Sinan; considers that the bulbs are connected in parallel.</p>
Q2	<p>You drop an apple from a certain height without speed and it hits the ground with a certain velocity. Assume that you cut the same apple in half and drop half of it at the same height. (Assume no air friction, please express your answers only verbally without writing formulas)</p> <p>a) What is the speed of the cut apple in half according to the speed of the whole apple?</p> <p>b) Explain the reason by considering the factors that affect its speed for the answer to option a.</p> <p>c) Explain the reason taking into account the energy conservation laws for your answer to option a.</p>	<p>a) 0 point: If student did not answer or gave incorrect answer (in this case the answers of b and c are ignored).</p> <p>1 point: Speed does not change. / Speeds are the same.</p> <p>b) 0 point: If student did not answer or gave incorrect answer.</p> <p>2 points (if option a is correct): only if height is specified in option b,</p> <p>2 points (if option a is correct): only if mass is specified in option b,</p> <p>3 points (if option a is correct): In option b, if both height and mass are specified:</p> <p>The objects that are allowed to fall free from the same height have different masses, but their velocities are the same. Because the velocity of an object that makes free fall movement is related to height but it is independent of mass.</p> <p>c) 0 point: If student did not answer or gave incorrect answer.</p> <p>4 points (if option a and b are correct): According to energy conservation laws, if an object has only potential energy when it is at a certain height and the object is allowed to fall free from the height, the height decreases as it falls to the ground as soon as it is allowed to fall but the object starts to accelerate.</p> <p>In this case, the kinetic energy increases while the potential energy decreases.</p> <p>As soon as the object hits the ground, it has only kinetic energy increased and the total mechanical energy is preserved.</p>



Question ID	Question Text	Scoring Key
Q3	What is Mechanical Energy? Please explain. (Please express your answers only verbally without writing formulas)	0 point: If student did not answer or gave incorrect answer. 1 point: Mechanical energy is the sum of potential and kinetic energy for conservative systems.
Q4	What does scientific Work mean? Please explain. (Please express your answers only verbally without writing formulas)	0 point: If student did not answer or gave incorrect answer. 1 point: Work is applying force to an object in a certain direction and moving that object in the direction of the applied force.

## From addiction to pervasiveness: Validation of the Smartphone Pervasiveness Scale in Turkish adolescents

Osman Urfa <sup>1\*</sup>, Recep Görgülü <sup>2</sup>

<sup>1</sup>Ministry of National Education, Burdur, Türkiye

<sup>2</sup>Bursa Uludağ University, Faculty of Sport Sciences, Psychology of Elite Performance Laboratory (PePLaB), Bursa, Türkiye

### ARTICLE HISTORY

Received: Nov. 5, 2024

Accepted: June 22, 2025

### Keywords:

Smartphone  
pervasiveness,  
Problematic smartphone  
use,  
Smartphone addiction,  
Validity,  
Reliability.

**Abstract:** The present study aimed to adapt the Smartphone Pervasiveness Scale (SPS) into Turkish and to examine its psychometric properties among Turkish adolescents. To this end, two studies were conducted. Study 1 explored the factor structure of the SPS through exploratory factor analysis (EFA), using data collected from 216 adolescents ( $M_{age} = 14.50$ ,  $SD = 1.55$ ). Study 2 employed confirmatory factor analysis (CFA) with multi-group analysis (MGA) on a separate sample of 314 adolescents ( $M_{age} = 13.87$ ,  $SD = 2.10$ ) to confirm the factor structure of the SPS and to assess measurement invariance across gender. In addition, Study 2 examined the associations between SPS scores and several external variables—problematic smartphone use, well-being, loneliness, psychological distress, and academic performance—as evidence of criterion-related validity. In both studies, Cronbach's alpha and composite reliability (CR) coefficients were calculated to assess reliability. EFA results in Study 1 supported a 7-item, single-factor structure, with factor loadings ranging from .46 to .67. CFA results in Study 2 confirmed this structure. Measurement invariance across gender was supported by the MGA. Moreover, criterion-related validity was demonstrated in Study 2: SPS scores were positively correlated with problematic smartphone use, loneliness, and psychological distress, and negatively correlated with well-being and academic performance. In both studies, Cronbach's alpha and CR coefficients were observed at .71 or higher. In conclusion, the Turkish version of the SPS is a valid and reliable instrument for assessing the pervasiveness of smartphone use among Turkish adolescents.

## 1. INTRODUCTION

Among adolescents, internet and smartphone use is growing every day, and smartphones have become one of the most popular (i.e., easy to reach, handy, availability of various options) tools for accessing the Internet (Mascheroni & Ólafsson, 2016). Many activities, such as using social media, sharing and viewing photos and videos, communicating with friends, following celebrities or popular people globally, playing online games, meeting new people, listening to music, and using it for educational purposes, are carried out via smartphones. So smartphones have a wide range of uses that can appeal to people of all ages. Although smartphones have

\*CONTACT: Osman URFA ✉ [dr.osmanurf@yahoo.com](mailto:dr.osmanurf@yahoo.com) 📍 Ministry of National Education, Burdur, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

many features that make our lives easier, they can also have many negative effects. Especially, childhood and adolescence are important developmental periods for physical, psychological, and physiological development (Steinberg, 2022), and studies show that internet and smartphone use among children is increasing year by year (TÜİK, 2024). Therefore, excessive and uncontrolled use of smartphones, such as continuous gaming or watching videos, may harm children's development.

Among the negative effects of smartphone use, it is considered one of the well-documented addictions in literature. The concept of smartphone addiction is based on the substance dependence and abuse diagnosis criteria in the DSM-IV. According to the DSM-IV, the criteria for substance dependence include an increased tolerance to the substance, withdrawal symptoms, a loss of control over substance use, and negative consequences of substance use on social, occupational, or recreational activities (American Psychiatric Association, 2000). Based on these criteria, smartphone addiction can be explained as follows: people lose control over their smartphone use and experience functional impairment, such as sleep disturbance. They also experience withdrawal symptoms (dysphoric, anxious mood, etc.), and want to use their smartphones more and more over time (Lin *et al.*, 2017). In addition, smartphone addiction is considered a behavioral addiction (Lin *et al.*, 2017), and behavioral addictions refer to the excessive, uncontrolled, and compulsive use of an activity without a chemical substance (Clark & Limbrick-Oldfield, 2013).

The existing psychological tests in the literature were developed following the DSM-IV. In this regard, the Smartphone Addiction Scale (SAS; Kwon *et al.*, 2013a) and its short form (Kwon *et al.*, 2013b) are two of the most commonly used scales. The concept of smartphone addiction is frequently used in the literature due to the effect of scales such as SAS. However, there is also criticism of the use of the term 'addiction'. For instance, Panova and Carbonell (2018) state that studies and definitions of smartphone addiction are inadequate in terms of addiction criteria, and these definitions can be more accurately described as overuse or problematic use. Therefore, many studies use the term 'problematic smartphone use' rather than 'smartphone addiction'. Nevertheless, it is still included as both 'addiction' (e.g., Olson *et al.*, 2022) and 'problematic use' (e.g., Sohn *et al.*, 2019; Urfa, 2024) in the literature. Studies on this subject have found a significant relationship between problematic smartphone use with high levels of depression, anxiety (Boumosleh & Jaalouk, 2017), psychological disturbance, emotion regulation difficulties (Squires *et al.*, 2021), frustration intolerance (Urfa, 2024), and low academic performance (Samaha & Hawi, 2016). Many systematic reviews and meta-analyses have also been conducted in this field of research (e.g., Sohn *et al.*, 2019; Shahidin *et al.*, 2022). For example, one of these aforementioned studies examined the relationship between problematic smartphone use and mental health. The research analyzed 41 studies published between 2011 and 2017 and found an association between problematic smartphone use and high levels of depression, anxiety, stress, and poor sleep quality (Sohn *et al.*, 2019). Accordingly, another study examined the relationship between problematic smartphone use and emotion dysregulation and found a moderate positive relationship (Shahidin *et al.*, 2022).

Many studies have been, and continue to be, conducted under the frameworks of smartphone addiction or problematic smartphone use. More recently, however, there has been a growing body of research that conceptualizes smartphone use not solely in terms of addiction or problematic use, but also in terms of pervasiveness (e.g., Chakraborty *et al.*, 2024). In other words, rather than categorizing individuals as either addicted or not addicted in a binary or descriptive manner, recent studies have begun to adopt a more analytical approach by examining how frequently individuals use smartphones and in which domains of their daily lives (Gerosa *et al.*, 2022). For instance, Gui and Gerosa (2021) developed the Smartphone Pervasiveness Scale (SPS) specifically to measure the frequency of smartphone use across different domains of daily life. The SPS assesses how frequently adolescents use smartphones across various domains of daily life (eating, school, watching films, etc.). One of the scale's

distinctive features is its emphasis on the social dimensions of smartphone use in everyday contexts, rather than the pathological aspects (Gerosa *et al.*, 2022). In other words, smartphone use is evaluated concerning its social and physiological roles, for example, during studying, spending time with family, or sleeping. Importantly, an individual may not meet criteria for addiction; yet, frequent smartphone use may still disrupt daily functioning. In this regard, the SPS makes a valuable contribution to the existing literature by broadening the conceptualization of smartphone use beyond strictly problematic or addictive frameworks.

The SPS is a relatively recent instrument, and while it holds significant potential for advancing research in this area, empirical studies employing the scale remain limited. Existing findings suggest that higher levels of smartphone pervasiveness are associated with increased time spent on smartphones (Chakraborty *et al.*, 2024) and lower academic performance among adolescents (Gerosa *et al.*, 2022). Although there are many studies on smartphone addiction and problematic use, there is limited literature on the pervasiveness of smartphones and how often they are used in which aspects of life. To address this limitation, further research using the SPS should be conducted, and cross-cultural comparisons should be made by adapting the SPS to different languages and cultures by examining its psychometric properties.

In the current study, the Turkish form of the SPS was created, and the validity and reliability studies of the Turkish form were examined. As in many cultures, smartphones are widely used from an early age in Türkiye. In 2024, the Turkish Statistical Institute (TURKSTAT) surveyed the use of information technologies among children. The results showed that 91.3% of children aged 6–15 use the internet, primarily for watching videos. The survey also revealed that 53.5% of 6–10-year-olds and 86.2% of 11–15-year-olds use smartphones (TÜİK, 2024). The current study examines the pervasiveness of smartphone use, and the TURKSTAT report shows that children in Türkiye start using smartphones at an early age. Additionally, the problematic use of smartphones has been examined among various groups, including adolescents, university students, and adults, in Türkiye. Studies have demonstrated that problematic smartphone use is associated with physiological and psychological constructs such as depression, anxiety, stress (Kabadayı, 2024), bedtime procrastination, sleep quality (Bozkurt *et al.*, 2024), frustration intolerance (Urfa, 2024), psychiatric symptoms, and emotion regulation difficulties (Gül *et al.*, 2019) in Turkish culture. Based on all of this, it is thought that adapting SPS to Turkish culture will generate a wealth of new studies and take smartphone research in a different direction. Studies will address smartphones in terms of not only addiction or problematic use, but also frequency of use. Thus, it will be possible to examine the consequences of adolescents using their smartphones more frequently in different areas of their lives.

The present research was conducted in two consecutive studies. In Study 1, the items of the SPS were translated into Turkish using the back-translation method, and the factor structure was examined through exploratory factor analysis (EFA). In Study 2, the factor structure identified in Study 1 was tested using confirmatory factor analysis (CFA) on an independent sample. Following the CFA, the measurement invariance of the SPS across gender was examined. Previous research conducted with children in Türkiye has shown that males tend to use the internet and smartphones more frequently than females and are more likely to engage in digital gaming (TÜİK, 2024). Therefore, consistent with the approach taken in the original validation study (Gerosa *et al.*, 2022), measurement invariance across gender was assessed using multi-group analysis (MGA).

Study 2 also examined the criterion-related validity of the SPS by investigating its associations with problematic smartphone use, psychological distress, well-being, loneliness, and academic performance. Prior research has shown that problematic smartphone use is associated with lower well-being, increased loneliness, greater psychological distress, and poorer academic outcomes (Mahapatra, 2019; van der Schuur *et al.*, 2015). Accordingly, in the present study, it was hypothesized that smartphone pervasiveness would be positively associated with

problematic smartphone use, loneliness, and psychological distress, and negatively associated with well-being and academic performance. Evidence of these expected relationships would provide support for the criterion-related validity of the SPS.

## 2. METHOD

### 2.1. Participants

Two different samples were recruited across the two studies in the present research. The first sample was used for exploratory factor analysis (EFA), while the second sample was employed for confirmatory factor analysis (CFA) and criterion-related validity testing. According to the literature, sample size guidelines for factor analysis recommend either a minimum of 10 participants per item or at least 200 participants (DeVellis, 2017). Given that the SPS consists of 7 items, a sample size of 70 participants would be the minimum; however, to ensure more robust and reliable factor analysis results, the target was to recruit at least 200 participants for each study. Since criterion-related validity was assessed in Study 2, sample size was further calculated with G\*Power 3.1.9.7 (Faul *et al.*, 2007). Assuming a medium effect size (0.3), an alpha level of .01, and statistical power of .99, the required minimum sample size was to be at least 222 participants. Accordingly, the recruitment target for Study 2 set at 222 or more participants.

Participants were secondary and high school students continuing their education from Burdur province, in Türkiye. A convenience sampling method was employed to recruit the participants. Permission to conduct the research was obtained from the Ministry of National Education for three high schools and three middle schools located in different neighborhoods of Burdur city center. The study procedures were first explained to school administrators, parents, and students. Only students whose parents and who themselves provided informed consent were invited to complete the study scales.

Descriptive statistics of the participants are presented in Table 1. In total, 530 secondary and high school students participated across the two studies, with 216 in Study 1 (Sample 1) and 314 in Study 2 (Sample 2). In Study 1, 31.48% of the participants were middle school students and 68.52% were high school students; 61.57% were boys and 38.43% were girls. In Study 2, 55.73% of the participants were middle school students and 44.27% were high school students; 44.59% were boys and 55.41% were girls. All participants reported using smartphones. The average age was 14.50 ( $SD = 1.55$ ) in Study 1 and 13.87 ( $SD = 2.10$ ) in Study 2.

**Table 1.** Participants.

		Sample 1		Sample 2	
		<i>f</i>	%	<i>f</i>	%
School	Middle school	68	31.48	175	55.73
	High school	148	68.52	139	44.27
Gender	Boys	133	61.57	140	44.59
	Girl	83	38.43	174	55.41
Devices in use*	Smartphone	216	100	314	100
	Tablet	48	22.22	91	28.98
	Computer	110	50.93	142	45.22
		Mean $\pm$ SD		Mean $\pm$ SD	
Age		14.50 $\pm$ 1.55		13.87 $\pm$ 2.10	
Sibling count		2.39 $\pm$ 0.83		2.18 $\pm$ 0.92	
Daily smartphone use (hours)		3.55 $\pm$ 2.32		3.50 $\pm$ 2.30	



## 2.2. Measures

### 2.2.1. Personal information sheet

The personal information sheet includes information about the participants' age, gender, grade level at school, daily smartphone usage time, and last academic year's grade points average (GPA).

### 2.2.2. Smartphone pervasiveness scale

The Smartphone Pervasiveness Scale was developed by Gui and Gerosa (2021) to measure how often adolescents use smartphones in different areas of life (eating, friends, school, etc.). The original scale is a 4-point Likert-type scale consisting of 5 items. Factor analysis confirmed the single-factor structure of the scale (RMSEA = .055, CFI = .991, TLI = .981) and the internal consistency coefficient is within acceptable limits (Cronbach alpha = .723). Later, two more items were added to the scale, and the 7-item form of the scale was organized according to a 5-point Likert-type response (Gui *et al.*, 2018). Gerosa *et al.* (2022) examined the validity and reliability of the scale in Italian adolescents. In this study, in which both exploratory and confirmatory factor analyses were performed, the 7-item and single-factor structure of the scale was confirmed.

### 2.2.3. General health questionnaire

The General Health Questionnaire (GHQ) was developed by Goldberg (1972, 1978) to measure psychological distress. The GHQ has 4 forms with 12, 28, 30 items. While the 12-item form measures psychological distress as a single dimension, the other forms analyze psychological distress according to subgroups. The 12 and 28-item forms of the inventory were adapted into Turkish by Kılıç (1996). The Turkish adaptation of the GHQ was administered to 121 patients referred to a psychiatric outpatient clinic. In 12-item Turkish form, the sensitivity and specificity of the Turkish form of the GHQ were calculated as .74 and .82, respectively. Cronbach's alpha internal consistency of the GHQ is .78, and test-retest reliability is .84 (Kılıç, 1996). The validity and reliability of the WHO-5 were examined in the current sample. For the construct validity, 12-item and single-factor structure of the GHQ was tested with CFA and acceptable fit values were obtained ( $S-B\chi^2 / df = 2.845$ , CFI = .917, TLI = .900, RMSEA = .077, SRMR = .050). Item factor loadings were found to be between .426 and .754. In addition, the Cronbach's alpha internal consistency coefficient of the questionnaire was .89 in the current study.

### 2.2.4. WHO 5 - Well-being index

WHO-5 (The World Health Organization Well-Being Index) is an inventory created by the World Health Organization to measure well-being (WHO, 1998). There are 5 items measuring well-being in the inventory, and these items are evaluated according to a 6-point Likert scale. Turkish adaptation of the inventory was conducted by Eser *et al.* (2019). In the Turkish adaptation study of the WHO-5, the EFA results showed that the 5-item, single-factor structure explained 58.5% of the total variance for adults and 63.9% for older adults. Item factor loadings ranged from .50 to .85. The CFA results showed that although the 5-factor structure was close to acceptable levels for adults (CFI = .989, NFI = .986, GFI = .987, RMSEA = .073, SRMR = .021) and older adults (CFI = .956, NFI = .954, GFI = .946, RMSEA = .166, SRMR = .043), it provided a better model fit for adults. The internal consistency coefficient calculated for the reliability of the WHO-5 was found to be .81 for adults and .86 for the elderly (Eser *et al.*, 2019). The validity and reliability of the WHO-5 were examined in the current sample. For the construct validity, the 5-item and single-factor structure of the WHO-5 was tested with CFA and excellent fit values were obtained ( $S-B\chi^2 / df = 1.556$ , CFI = .991, TLI = .973, RMSEA = .066, SRMR = .021). Item factor loadings were found to be between .625 and .725. In addition, the internal consistency coefficient of the inventory was found to be .76 in the current study.

### 2.2.5. Smartphone addiction scale - Short version

It is the shortened version of the 33-item Smartphone Addiction Scale (SAS) developed by Kwon *et al.* (2013a) by Kwon *et al.* (2013b). The original 33-item SAS was subjected to content validity evaluation by seven expert reviewers, from which a condensed 10-item version (SAS-SV) was subsequently developed for administration to adolescent samples. The SAS-SV consists of 10 items and single dimension to measure smartphone addiction. The scale items are evaluated according to a 6-point Likert scale (1: strongly disagree, 6: strongly agree). The Cronbach alpha internal consistency coefficient is .91, and the corrected item total score correlations are between .50 and .80 (Kwon *et al.*, 2013b). The Turkish adaptation of the scale was conducted by Şata and Karip (2017) in Turkish adolescents. In this study, ten items and one dimension were found to be consistent with the original scale. As a result of the confirmatory factor analysis for the construct validity of the Turkish form, acceptable fit indices were obtained (GFI = .93, CFI = .99, RMSEA = .064). The internal consistency coefficients calculated for the reliability of the scale were also high (Cronbach's alpha = .90, McDonald's omega = .94) (Şata & Karip, 2017). The validity and reliability of the SAS-SV were examined in the current sample. For the construct validity, the 10-item and single-factor structure of the SAS-SV was tested with CFA, and excellent fit values were obtained ( $S-B\chi^2 / df = 1.209$ , CFI = .996, TLI = .994, RMSEA = .027, SRMR = .058). Item factor loadings were found to be between .393 and .802. In the current study, the Cronbach's alpha internal consistency coefficient of the SAS-SV was .86.

### 2.2.6. UCLA loneliness scale - Short form

The UCLA Loneliness Scale, one of the most widely used instruments for assessing loneliness, was employed in this study. The original scale comprises 20 items; however, a shortened version was developed by Hays and DiMatteo (1987). To create this abbreviated form, Hays and DiMatteo (1987) conducted an exploratory analysis (EFA) on data collected from 199 college students who completed all 20 items. After EFA, a short form consisting of eight items and a single factor emerged. The eight items explained 67.44% of the total variance, and item factor loadings ranged from .31 to .73. The Cronbach's alpha internal consistency coefficient of the 8 items is .84 (Hays & DiMatteo, 1987). Turkish adaptation of the scale was conducted by Yıldız and Duy (2014) in 293 high school students aged between 14 and 19. In the adaptation study, firstly, principal component analysis was used in conjunction with parallel analysis to determine the item-factor structure of the scale. In the analyses, one item was removed because its factor loading was less than .30, and the Turkish form of the scale showed a single-factor structure with seven items. Seven items explained 40.99% of the total variance, and item factor loadings ranged from .31 to .73. The 7-item, single-factor structure of the scale was confirmed by the CFA conducted after the EFA ( $\chi^2 / df = 1.94$ , RMSEA = .06, SRMR = .04, GFI = .97, AGFI = .95, CFI = .98, NFI = .96, NNFI = .97). Cronbach's alpha internal consistency coefficient calculated for reliability is .74, and the composite reliability coefficient is .75 (Yıldız & Duy, 2014). The validity and reliability of the scale were examined in the current sample. For the construct validity, the 7-item and single-factor structure of the scale was tested with CFA, and acceptable fit values were obtained ( $S-B\chi^2 / df = 3.155$ , CFI = .944, TLI = .916, RMSEA = .075, SRMR = .075). Item factor loadings were found to be between .396 and .907. The internal consistency coefficient of the scale was found to be .73 in the current study.

## 2.3. Procedure

Permission for the Turkish adaptation of the Smartphone Pervasiveness Scale was initially obtained from Marco Gui via e-mail. Following permission, ethical approval was obtained from the Social and Human Sciences Research and Publication Ethics Committee of Bursa Uludağ University (date: 27.09.2024, number: 2024/9, decision number: 7). After approval from the ethics committee, permission to implement the data collection tools in secondary and high schools was obtained from the Turkish Ministry of National Education.

Following approval from the Ethics Committee and the Turkish Ministry of National Education, the scale items were translated into Turkish using the approach recommended by Beaton *et al.* (2000). Initially, two English teachers, both native Turkish speakers, independently translated items into Turkish. These translations were reviewed and synthesized by the researchers into a single version. Subsequently, two different English teachers performed a back-translation of this Turkish version into English. The original scale items and back-translated versions were then compared to identify and resolve any discrepancies or ambiguities. The iterative process resulted in the finalized Turkish version of the scale (see [Appendix 1](#) and [Appendix 2](#)). To assess the clarity and comprehensibility of the translated items, the Turkish form was administered to ten eighth-grade students and two psychological counselors using the think-aloud protocol (Jääskeläinen, 2012). This procedure allowed for the examination of participants' cognitive processes while completing the scale and provided qualitative feedback on the items. All participants reported a clear understanding of the items, and no negative feedback was received. After the Turkish translation of the scale items was completed, the schools were contacted. Schools were selected to represent different parts of the city and regions with different socio-demographic environments. Data collection was subsequently conducted in three secondary schools and three high schools. Before participation, both students and their parents were provided with detailed information about the study, and informed consent was obtained from both parties. Only students who provided both their own consent and parental consent participated in the study.

## 2.4. Data Analysis

The study investigated the validity and reliability of the SPS using data collected from two independent samples. Prior to data analysis, outliers were identified and removed from both datasets. Specifically, five participants from the first sample and twelve participants from the second sample were excluded based on Mahalanobis distance. Subsequent analyses were conducted on the finalized datasets.

Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were conducted to test the construct validity of the SPS. In the analysis of the data, exploratory factor analysis (EFA) was conducted first. The Kaiser-Meyer-Olkin (KMO) value for EFA was found to be .739, and Bartlett's sphericity test was significant ( $\chi^2 = 209.516$ ,  $df = 21$ ,  $p < .001$ ). These values indicate that the dataset is suitable for factor analysis (Field, 2012). Parallel analysis (Horn, 1965) was used to determine the number of factors as a result of EFA. As a rotation technique, oblimin rotation, which is one of the oblique rotation methods, was used because it was thought that there would be a relationship between usage frequencies in different areas of smartphones in daily life (Hair *et al.*, 2014). Furthermore, principal axis factoring was used as a factoring method in EFA.

The item-factor structure obtained after EFA was tested in another sample (sample 2) using confirmatory factor analysis (CFA). As the dataset did not fulfill the multivariate normality assumption (Mardia Skewness = 1376.11,  $p < .01$ , Mardia kurtosis = 21.87,  $p < .01$ ; Korkmaz *et al.*, 2014), robust goodness-of-fit indices with Satorra-Bentler (S-B) correction were used to analyze the data (Gana & Broc, 2019). For this purpose,  $\chi^2 / df$ , CFI, TLI, RMSEA and SRMR goodness-of-fit indices were used. Among these indices, CFI and TLI values above .90 and RMSEA and SRMR values below .08 indicate that the factor structure of the scale is acceptable (Hu & Bentler, 1999). Following the CFA, a multi-group analysis (MGA) was conducted to test the measurement invariance of SPS across gender (boys and girls). Boys were coded as '0' and girls as '1'. Changes in CFI ( $\Delta CFI$ ), RMSEA ( $\Delta RMSEA$ ), and SRMR ( $\Delta SRMR$ ) were tested for measurement invariance. Measurement invariance is indicated by  $\Delta CFI$ ,  $\Delta RMSEA$  and  $\Delta SRMR$  values less than |.01|, |.015|, and |.030|, respectively (Chen, 2007; Cheung & Rensvold, 2002). First, configural invariance, which tests whether the factor structure of the scale is the equal across subgroups (boys-girls), was examined. Then metric, strong, and strict

invariances were tested respectively. Metric invariance indicates factor loadings, strong invariance indicates factor loadings and intercepts, and strict invariance indicates that factor loadings, intercepts, and measurement error variance are equal across subgroups (Gana & Broc, 2019).

In order to test the criterion-related validity of the SPS, the relationship between the SPS scores with well-being, loneliness, psychological distress, problematic smartphone use, and academic performance was examined. Therefore, Pearson's product-moment correlation analysis was used to examine the relationship between variables. Finally, the Cronbach's alpha internal consistency and the composite reliability (CR) coefficients were analyzed for the reliability of the SPS. Both coefficients of .70 and above are known as acceptable reliability (Hair *et al.*, 2014). IBM SPSS Statistics 22 and JASP 0.18.3 were used for data analysis.

### 3. RESULTS

#### 3.1. Item Analysis (Study 1 and 2)

Table 2 shows mean, standard deviation, and corrected item total score correlation coefficients of the scale items. In Sample 1, the item means ranged from 1.26 (i6) to 2.88 (i7), and the corrected item total score correlations ranged from .32 (i6) to .48 (i2). In sample 2, the item means ranged from 1.21 (i6) to 2.90 (i7) and the corrected item total score correlations ranged from .38 (i1) to .59 (i4).

**Table 2.** Result of the item analysis.

	Sample 1 (Study 1)			Sample 2 (Study 2)		
	Mean	SD	Corrected item total score correlation	Mean	SD	Corrected item total score correlation
m1	1.69	1.16	.36	1.64	1.17	.38
m2	2.65	1.11	.48	2.42	1.20	.55
m3	2.13	1.14	.41	2.09	1.20	.45
m4	2.09	1.33	.45	2.18	1.44	.59
m5	2.66	1.52	.47	2.46	1.56	.56
m6	1.26	0.92	.32	1.21	.77	.43
m7	2.88	1.30	.42	2.90	1.35	.49

#### 3.2. Exploratory Factor Analysis and Reliability (Study 1)

Exploratory factor analysis (EFA) was conducted to explore the factor structure of the scale (Table 3).

**Table 3.** Factor loadings and reliability coefficients obtained after EFA and CFA.

Items	EFA (Study 1)			CFA (Study 2)		
	Factor loadings	$\alpha$	CR	Factor loadings	$\alpha$	CR
m1	.53	.71	.79	.42	.77	.77
m2	.66			.65		
m3	.59			.49		
m4	.63			.70		
m5	.67			.69		
m6	.46			.44		
m7	.60			.57		

The parallel analysis conducted for the EFA indicated that the SPS had a single factor structure. This single-factor structure explained 35% of the total variance of the scale. The factor loadings of the items were between .46 (m6) and .67 (m5). The Cronbach alpha internal consistency coefficient calculated for reliability was found to be .71, while the composite reliability coefficient was found to be .79.

### 3.3. Confirmatory Factor Analysis and Reliability (Study 2)

The 7 items and the single-factor structure obtained as a result of EFA were tested in sample 2 using confirmatory factor analysis (CFA) (Tables 3 and 4). Acceptable fit values were obtained after CFA ( $S-B\chi^2 = 41.959$ ,  $df = 14$ , CFI = .938, TLI = .908, RMSEA = .078, SRMR = .045) (Table 4). The factor loadings of the items ranged from .42 (i1) to .70 (i4). The Cronbach's alpha internal consistency coefficient calculated for reliability was .77, while the composite reliability coefficient was .77 (Table 3).

**Table 4.** CFA model fit indices and measurement invariance across gender.

	S- $B\chi^2$ (df)	TLI	CFI ( $\Delta$ CFI)	RMSEA ( $\Delta$ RMSEA)	SRMR ( $\Delta$ SRMR)
All participants	41.959 (14)	.908	.938	.078	.045
Boys	28.236 (14)	.898	.932	.075	.053
Girls	32.630 (14)	.895	.930	.077	.054
Configural invariance	60.867 (28)	.896	.931	.076	.057
Metrik invariance	73.367 (34)	.891	.926 (.005)	.086 (.010)	.068 (.011)
Scalar invariance	80.067 (40)	.889	.924 (.002)	.089 (.003)	.071 (.003)
Strict invariance	88.681 (47)	.893	.921 (.003)	.084 (.005)	.078 (.006)

### 3.4. Measurement Invariance of the SPS (Study 2)

Measurement invariance results across gender are shown in Table 4. The CFA results indicated that the 7-item and single-factor structure of the SPS had acceptable fit indices for both girls and boys. For measurement invariance, configural, metric, scalar, and strict invariance were tested respectively, and the results of the analyses showed that  $\Delta$ CFI was between .002 and .005,  $\Delta$ TLI was between .002 and .005,  $\Delta$ RMSEA was between .003 and .010, and  $\Delta$ SRMR was between .003 and .011.

### 3.5. Criterion-Related Validity (Study 2)

For criterion-related validity, the relationship between smartphone pervasiveness scores with problematic smartphone use, well-being, loneliness, psychological distress, and academic performance were examined using correlation analysis (Table 5).

**Table 5.** Descriptive statistics and correlation coefficients of the variables.

	Descriptives				Correlation coefficients				
	Mean	SD	Skewness	Kurtosis	SP	PSU	WHO 5	UCLA	PD
SP	2.13	0.82	1.17	1.57	-				
PSU	2.54	1.06	0.47	-0.40	.42**	-			
WHO 5	2.65	1.20	-0.11	-0.57	-.18**	-.27**	-		
UCLA	1.87	0.63	0.61	-0.25	.19**	.22**	-.26**	-	
PD	2.00	0.70	0.91	0.84	.31**	.40**	-.57**	.53**	-
GPA	83.08	12.22	-0.73	-0.22	-.35**	-.14*	.18**	-.14*	-.12

SP: Smartphone pervasiveness; PSU: Problematic smartphone use; WHO 5: WHO 5 - Well-being index; UCLA: UCLA loneliness; PD: Psychological distress; GPA: Last academic year's GPA.

\*\*  $p < .01$  \*  $p < .05$



The analysis revealed a positive relationship between smartphone pervasiveness with problematic smartphone use ( $r = .42, p < .01$ ), loneliness ( $r = .19, p < .01$ ), and psychological distress ( $r = .31, p < .01$ ). At the same time, smartphone pervasiveness was negatively correlated with well-being ( $r = -.18, p < .01$ ) and academic GPA ( $r = -.35, p < .01$ ).

#### 4. DISCUSSION and CONCLUSION

The present study aims to examine the psychometric properties of the Turkish version of the Smartphone Pervasiveness Scale (SPS). Accordingly, the Turkish form of the SPS was created and its validity and reliability analyses were examined in two different samples. Exploratory factor analysis (EFA) was conducted to investigate the factor structure of the scale in Sample 1. Subsequently, confirmatory factor analysis (CFA) was employed with Sample 2 to validate the factor structure and assess the criterion-related validity of the scale. Both samples underwent item analysis and reliability testing accordingly.

Item analysis included the examination of item means, standard deviations, and corrected item-total score correlations. In the current study, item means ranged from 1.26 (i6: 'At school, during lessons') to 2.88 (i7: 'While you are watching a movie or a TV show'), whereas in the original study (Gerosa *et al.*, 2022) they ranged from 1.6 (i1: 'At dinner with your family') to 2.7 (i5: 'First thing in the morning, when you wake up') among Italian adolescents. Although the overall mean scores were similar across cultures, the items with the lowest and highest mean scores differed. Turkish adolescents reported the lowest smartphone use during school lessons, whereas Italian adolescents reported the lowest use during family dinners. Conversely, Turkish adolescents indicated the highest usage while watching films or TV programs, while Italian adolescents reported the highest use immediately upon waking. Corrected item-total correlations were also analyzed, with the lowest correlations observed being .32 (item 6: 'At school, during lessons') in Sample 1 and .38 (item 1: 'At dinner with your family') in Sample 2. All correlation coefficients exceeded the recommended threshold of .30, indicating strong item discrimination (Büyüköztürk, 2012). The relatively low mean and correlation for item 6 are consistent with the common restriction on smartphone use in Turkish classrooms.

According to the EFA and CFA for construct validity, the 7-item and single-factor structure of the original scale (Gerosa *et al.*, 2022) was confirmed in the Turkish adolescents. EFA showed that the 7 items explained 35% of the total variance of the SPS. In the original form of the SPS (Gerosa *et al.*, 2022), similar results were found to the present study and explained 40% of the total variance of the scale. It is recommended that the explained variance ratio should be 66% (2/3) and above, but a value of 30% and above is considered acceptable for single-factor scales (Büyüköztürk, 2012). Therefore, it is a clear indication that the variance explained in the current study is sufficient. However, the scale could be revised in future research to increase the explained variance ratio. The scale currently consists of seven items and does not cover all the places/areas where adolescents can use their smartphones (e.g., during public transport/vehicle travel or while resting after an activity). While the scale in its present form is specifically designed for adolescent participants, future research endeavors may warrant the development of an adult version that incorporates contextually relevant items addressing adult smartphone use patterns, such as usage during professional activities, driving, and commercial transactions. Thus, it is thought that adding items for the possible usage areas of smartphones could increase the explained variance ratio. Additionally, the EFA results showed item factor loadings between .46 (i6: 'At school, during the lessons') and .67 (i5: 'First thing in the morning, when you wake up'). Similar to the current study, the factor loadings of the SPS in the original study (Gerosa *et al.*, 2022) were found to be between .45 (i7: 'While you are watching a movie or a TV show') and .63 (i3: 'While you are studying'). The original and Turkish forms of the SPS were found to have similar factor loadings in EFA. According to the literature, the item factor loadings should be at least .364 for a sample size of 200 for EFA (Field, 2012). Therefore, the item factor loadings were found to be sufficient.

The CFA results based on EFA have acceptable goodness of fit indices ( $S-B\chi^2 = 41.959$ ,  $df = 14$ ,  $CFI = .938$ ,  $TLI = .908$ ,  $RMSEA = .078$ ,  $SRMR = .045$ ) (Kline, 2011). In the CFA results, the lowest item factor loading was found to be .42. For CFA, the threshold value of .40 was determined as the lowest factor loading (Hair et al., 2017). Therefore, all items have sufficient factor loadings. Following the CFA, the measurement invariance of the SPS was analysed according to the child's gender using a multi-group analysis (MGA). As a result of the MGA,  $\Delta CFI$  values were less than  $|.01|$ ,  $\Delta RMSEA$  values were less than  $|.015|$  and  $\Delta SRMR$  values were less than  $|.030|$ ; therefore, it was seen that the SPS provided strict invariance. In other words, the factor loadings, intercepts, and measurement error variance of the SPS Turkish form are equal across gender (Chen, 2007; Cheung & Rensvold, 2002). In line with the current study, the original study of the SPS (Gerosa et al., 2022) showed measurement invariance across children's gender. Furthermore, Gerosa et al. (2022) demonstrated that the SPS exhibited measurement invariance across different levels of parental education. Consequently, it is recommended that future research investigate the measurement invariance of the Turkish version of the SPS concerning parents' educational attainment.

Cronbach's alpha internal consistency coefficient and composite reliability coefficient were calculated for the reliability of the SPS. These coefficients were found to be .71 and .79 in Study 1 and .77 and .77 in Study 2, respectively. Both coefficients of .70 and above are considered acceptable reliability (Hair et al., 2014). Therefore, the SPS was found to have an acceptable level of internal consistency. In the original study of the SPSS (Gerosa et al., 2022), the Cronbach's alpha internal consistency coefficient was also found to be .73, which is similar to the current study.

The criterion-related validity of the SPS was assessed by examining its relationships with problematic smartphone use, well-being, loneliness, psychological distress, and academic GPA. Consistent with the hypotheses, smartphone pervasiveness was significantly positively correlated with problematic smartphone use, loneliness, and psychological distress, and significantly negatively correlated with well-being and academic GPA. These findings provide further support for the validity of the SPS. The scale aims to evaluate the impact of pervasive smartphone use on individuals' physiological, psychological, and social functioning (Gerosa et al., 2022), and the results of the correlation analyses align with this objective.

Pervasive smartphone use was found to be associated with adverse social, psychological, and academic outcomes, including increased loneliness, elevated psychological distress, and reduced academic performance. Thus, pervasive smartphone use appears to pose risks to the psychological, social, and academic functioning of Turkish adolescents. In light of these findings, it is recommended that efforts be made to reduce the frequency of smartphone use to support adolescent mental health. Specifically, ministries and municipalities, in collaboration with institutions such as schools and youth centers, should provide school-based psychological services, as well as opportunities for engagement in sports, arts, and social activities from an early age.

The results of the correlation analysis reveal an additional significant finding that supports the conceptual foundation of the SPS. Specifically, the correlation coefficients between problematic smartphone use and well-being, loneliness, and psychological distress are higher than those between smartphone pervasiveness and these variables. Conversely, the correlation between academic performance and smartphone pervasiveness is stronger than that between academic performance and problematic smartphone use. This suggests that problematic smartphone use is more closely associated with well-being, loneliness, and psychological distress, whereas smartphone pervasiveness is more strongly linked to academic performance. The SPS primarily focuses on the widespread and excessive use of smartphones, rather than on pathological behaviors such as addiction. Consequently, it emphasizes how individuals engage with family, peers, and academic responsibilities (Gerosa et al., 2022). Therefore, the stronger association between SPS scores and academic performance, relative to psychological variables,

is an anticipated outcome. Future research investigating the SPS within the context of social relationships—such as adolescents’ socialization patterns with family and peers and their communication styles—would further contribute to understanding these dynamics.

Based on the findings from all analyses, it can be concluded that the Turkish version of the SPS is a valid and reliable instrument for assessing the pervasiveness of smartphone use among adolescents. While numerous studies in the literature focus on constructs such as 'smartphone addiction' or 'problematic smartphone use,' the concept of 'smartphone pervasiveness' is relatively novel and offers a fresh perspective for research in this area. This study explored the relationships between smartphone pervasiveness and problematic smartphone use, well-being, loneliness, psychological distress, and academic GPA, yielding results consistent with theoretical expectations.

There remain many variables that warrant further investigation. Future research could explore the associations between smartphone pervasiveness and adolescents’ social relationships (e.g., family and peer interactions), physical and mental health outcomes, and psychiatric diagnoses. Additionally, the SPS may serve as a useful measure to evaluate the effectiveness of experimental interventions aimed at reducing problematic smartphone use by examining whether such interventions decrease the pervasiveness of smartphone use among adolescents. Previous longitudinal research has highlighted the association between parental addiction and child addiction to smartphones (Jeong *et al.*, 2024). Thus, it is recommended that future studies examine the relationship between parental and adolescent smartphone pervasiveness to inform family-based interventions. Moreover, comparative studies assessing the effects of smartphone usage frequency across different age groups and socioeconomic statuses would provide valuable insights. Further psychometric research is also encouraged. Large-scale normative studies using item response theory could deepen understanding of the SPS’s measurement properties. While the current study established internal consistency reliability, future work should assess test-retest reliability to confirm the scale’s temporal stability.

Finally, several limitations of the present study should be acknowledged. The primary limitation concerns the sample, which was drawn from a single province in Türkiye, a country characterized by considerable ethnic and cultural diversity. Additionally, all measures employed were self-reported, which may introduce response biases. Furthermore, the cross-sectional design, with data collected at a single time point, precludes any conclusions about causal relationships among the variables. Future longitudinal studies examining the outcomes and stability of the SPS over time are recommended to address these limitations and contribute further to the literature.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Bursa Uludağ University, Social and Human Sciences Research and Publication Ethics Committee, 2024/9, decision number: 7.

### Contribution of Authors

**Osman Urfa:** Conceptualization, Investigation, Resources, Software, Formal Analysis, and Writing-original draft. **Recep Görgülü:** Conceptualization, Writing, Review, and Editing.

### Orcid

Osman Urfa  <https://orcid.org/0000-0002-9821-671X>

Recep Görgülü  <https://orcid.org/0000-0003-2590-4893>

### REFERENCES

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. American Psychiatric Association.

- Beaton, D.E., Bombardier, C., Guillemin, F., & Ferraz, M.B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186-3191. <https://doi.org/10.1097/00007632-200012150-00014>
- Bozkurt, A., Demirdöğen, E.Y., & Akıncı, M.A. (2024). The association between bedtime procrastination, sleep quality, and problematic smartphone use in adolescents: A mediation analysis. *The Eurasian Journal of Medicine*, 56(1), 69-75. <https://doi.org/10.5152/eurasianjmed.2024.23379>
- Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı [Handbook of data analysis for the social sciences]*. Pegem Akademi.
- Cai, Z., Mao, P., Wang, Z., Wang, D., He, J., & Fan, X. (2023). Associations between problematic internet use and mental health outcomes of students: A meta-analytic review. *Adolescent Research Review*, 8(1), 45–62. <https://doi.org/10.1007/s40894-022-00201-9>
- Chakraborty, S., Gui, M., Gerosa, T., & Marciano, L. (2024). Testing the validity of the smartphone pervasiveness scale for adolescents with self-reported objective smartphone use data. *Digital Health*, 10. <https://doi.org/10.1177/20552076241234744>
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464-504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- Clark, L., & Limbrick-Oldfield, E.H. (2013). Disordered gambling: a behavioral addiction. *Current Opinion in Neurobiology*, 23, 655-659. <http://dx.doi.org/10.1016/j.conb.2013.01.004>
- DeVellis, R.F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.
- Eser, E., Çevik, C., Baydur, H., Güneş, S., Esgin, T.A., Öztekin, Ç., Eker, E., Gümüşsoy, U., Eser, G.B., & Özyurt, B. (2019). Reliability and validity of the Turkish version of the WHO-5, in adults and older adults for its use in primary care settings. *Primary Health Care Research ve Development*, 20(e100), 1-7. <https://doi.org/10.1017/S1463423619000343>
- Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. <https://doi.org/10.3758/bf03193146>
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.
- Gana, K., & Broc, G. (2019). *Structural equation modeling with lavaan*. John Wiley & Sons.
- Gerosa, T., Gui, M., & Büchi, M. (2022). Smartphone use and academic performance: a pervasiveness approach beyond addiction. *Social Science Computer Review*, 40(6), 1542–1561. <https://doi.org/10.1177/08944393211018969>
- Goldberg, D.P. (1972). *Detecting psychiatric illness by questionnaire*. Oxford University Press.
- Goldberg, D.P. (1978). *Manual of the General Health Questionnaire*. NFER-Nelson.
- Gui, M., & Gerosa, T. (2021). Smartphone pervasiveness in youth daily life as a new form of digital inequality. In E. Hargittai (Ed.), *The handbook of digital inequality* (pp. 131–147). Edward Elgar Publishing.
- Gui, M., Gerosa, T., Garavaglia, A., Petti, L., & Fasoli, M. (2018). *Digital well-being. Validation of a digital media education programme in high schools*. Report. Research Center on Quality of Life in the Digital Society. University of Milano Bicocca.
- Gül, H., Fırat, S., Sertçelik, M., Gül, A., Gürel, Y., & Kılıç, B.G. (2019). Cyberbullying among a clinical adolescent sample in Turkey: effects of problematic smartphone use, psychiatric



- symptoms, and emotion regulation difficulties. *Psychiatry and Clinical Psychopharmacology*, 29(4), 547–557. <https://doi.org/10.1080/24750573.2018.1472923>
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2014). *Multivariate data analysis*. Pearson.
- Hair, J.F., Hult, G.T.M., Ringle, C.M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM)*. SAGE Publications.
- Hays, R.D., & DiMatteo, M.R. (1987). A short-form measure of loneliness. *Journal of Personality Assessment*, 51, 69–81. [https://doi.org/10.1207/s15327752jpa5101\\_6](https://doi.org/10.1207/s15327752jpa5101_6)
- Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.
- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Jääskeläinen, R. (2012). Think-aloud protocol. In Y. Gambier & L. Doorslaer (Eds.), *Handbook of translation studies: Volume 1* (pp. 371–373). John Benjamins.
- Jeong, K.H., Kim, S., Ryu, J.H., & Lee, S. (2024). A longitudinal relationship between mother's smartphone addiction to child's smartphone addiction. *International Journal of Mental Health and Addiction*, 22(4), 1771–1782. <https://doi.org/10.1007/s11469-022-00957-0>
- Kabadayı, F. (2024). The examination of the relationship between irrational beliefs, depression, anxiety, stress and internet addiction in emerging adults. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi*, 14(3), 1645-1667. <https://doi.org/10.30783/nevsosbilen.1514229>
- Kılıç, C. (1996). Genel Sağlık Anketi: Geçerlik ve güvenirlik çalışması [General Health Questionnaire: Validity and reliability study]. *Türk Psikiyatri Dergisi*, 7(1), 3–11.
- Kline, R.B. (2011). *Principles and practice of structural equation modeling*. Guilford Publications.
- Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, 6(2), 151–62.
- Kwon, M., Kim, D.J., Cho, H., & Yang, S. (2013b). The smartphone addiction scale: Development and validation of a short version for adolescents. *PLoS ONE*, 8(12). <https://doi.org/10.1371/journal.pone.0083558>
- Kwon, M., Lee, J.Y., Won, W.Y., Park, J.W., Min, J.A., Hahn, C., Gu, X., Choi, J.H., & Kim, D.J. (2013a). Development and validation of a Smartphone Addiction Scale (SAS). *PLoS ONE*, 8(2). <https://doi.org/10.1371/journal.pone.0056936>
- Lin, Y.H., Lin, S.H., Yang, C.C.H., & Kuo, T.B.J (2017). Psychopathology of everyday life in the 21st century: Smartphone addiction. In C. Montag (Ed.), *Internet addiction: neuroscientific approaches and therapeutical implications including smartphone addiction* (pp. 339–358). Springer.
- Mahapatra, S. (2019). Smartphone addiction and associated consequences: Role of loneliness and self-regulation. *Behaviour & Information Technology*, 38(8), 833-844. <https://doi.org/10.1080/0144929X.2018.1560499>
- Mascheroni, G., & Olafsson, K. (2016). The mobile Internet: Access, use, opportunities and divides among European children. *New Media & Society*, 18(8), 1657-1679. <https://doi.org/10.1177/14614448145679>
- Olson, J.A., Sandra, D.A., Colucci, E.S., Bikaii, A.A., Chmoulevitch, D., Nahas, J., Raz, A., & Veissiere, S.P.L. (2022). Smartphone addiction is increasing across the world: A meta-analysis of 24 countries. *Computers in Human Behavior*, 129. <https://doi.org/10.1016/j.chb.2021.107138>
- Panova, T., & Carbonell, X. (2018). Is smartphone addiction really an addiction? *Journal of Behavioral Addictions*, 7(2), 252–259. <https://doi.org/10.1556/2006.7.2018.49>



- Samaha, M., & Hawi, N.S. (2016). Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. *Computers in Human Behavior*, 57, 321–325. <https://doi.org/10.1016/j.chb.2015.12.045>
- Shahidin, S.H., Midin, M., Sidi, H., Choy, C.L., Nik Jaafar, N.R., Mohd Salleh Sahimi, H., & Che Roos, N.A. (2022). The relationship between emotion regulation (ER) and problematic smartphone use (PSU): A systematic review and meta-analyses. *International Journal of Environmental Research and Public Health*, 19(23). <https://doi.org/10.3390/ijerph192315848>
- Sohn, S.Y., Rees, P., Wildridge, B., Kalk, N.J., & Carter, B. (2019). Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: a systematic review, meta-analysis and GRADE of the evidence. *BMC Psychiatry*, 19(356). <https://doi.org/10.1186/s12888-019-2350-x>
- Squires, L.R., Hollett, K.B., Hesson, J., & Harris, N. (2021). Psychological distress, emotion dysregulation, and coping behaviour: A theoretical perspective of problematic smartphone use. *International Journal of Mental Health and Addiction*, 19(4), 1284–1299. <https://doi.org/10.1007/s11469-020-00224-0>
- Steinberg, L. (2022). *Adolescence*. McGraw Hill.
- Şata, M., & Karip, F. (2017). Turkish culture adaptation of Smartphone Addiction Scale-short version for adolescents. *Cumhuriyet International Journal of Education*, 6(4), 426–440. <https://doi.org/10.30703/cije.346614>
- TÜİK (2024). Çocuklarda bilişim teknolojileri kullanım araştırması, 2024 [Information technology usage survey in children, 2024]. <https://data.tuik.gov.tr/Bulten/Index?p=Cocuklarda-Bilisim-Teknolojileri-Kullanim-Arastirmasi-2024-53638>
- Urfa, O. (2024). A conditional process model to explain problematic smartphone use: The interaction among frustration intolerance, duration of use, and gender. *Psihologija*, 57(2), 215–226. <https://doi.org/10.2298/PSI220627017U>
- van der Schuur, W.A., Baumgartner, S.E., Sumter, S.R., & Valkenburg, P.M. (2015). The consequences of media multitasking for youth: A review. *Computers in Human Behavior*, 53, 204–215. <http://dx.doi.org/10.1016/j.chb.2015.06.035>
- WHO (1998). *Wellbeing measures in primary health care/the depcare project*. WHO Regional Office for Europe: Copenhagen.
- Yıldız, M.A., & Duy, B. (2014). Adaptation of the short-form of the UCLA Loneliness Scale (ULS-8) to Turkish for the adolescents. *Düşünen Adam the Journal of Psychiatry and Neurological Sciences*, 27, 194–203. <https://doi.org/10.5350/DAJPN2014270302>

## APPENDICES

### **Appendix 1.** *English items of the SPS.*

- i1 - At dinner with your family
- i2 - while you are spending time with your friends
- i3 - while you are studying
- i4 - during the night, if you wake up
- i5 - First thing in the morning, when you wake up
- i6 - At school, during the lessons
- i7 - While you are watching a movie or a TV show

### **Appendix 2.** *Turkish items of the SPS.*

- i1 - Ailenizle akşam yemeği yerken
- i2 - Arkadaşlarınızla vakit geçirirken
- i3 - Ders çalışırken
- i4 - Gece, uyandığınızda
- i5 - Sabahın ilk saatlerinde, uyandığınızda
- i6 - Okulda, ders sırasında
- i7 - Bir film veya TV programı izlerken

## Exploring trends in psychometrics literature through a structural topic model

Kübra Atalay Kabasakal<sup>1\*</sup>, Duygu Koçak<sup>2</sup>, Rabia Akcan<sup>3</sup>

<sup>1</sup>Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

<sup>2</sup>Alanya Alaaddin Keykubat University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye

<sup>3</sup>Republic of Turkey Ministry of National Education, Afyonkarahisar, Türkiye

### ARTICLE HISTORY

Received: Mar. 7, 2025

Accepted: July 8, 2025

### Keywords:

Structural topic modelling,

Psychometrics,

Trend analysis,

Latent dirichlet allocation,

Text mining.

**Abstract:** The digitalization of knowledge has made it increasingly challenging to find and discover relevant information, leading to the development of computational tools to assist in organizing, searching, and comprehending vast amounts of information. In fields like psychometrics, which involve large datasets, a comprehensive examination of research trends, as well as understanding the prominence of various themes and their evolution over time through these tools, is essential for assessing the dynamic structure of the field. This study aims to explore the themes addressed in publications from eleven leading journals in psychometrics and to determine the overall distribution of topics. To achieve this, structural topic modelling has been employed. A comprehensive analysis of 8,523 article abstracts sourced from the Web of Science database revealed the existence of fourteen topics within the publications. “Scale Development and Validation” emerged as the most prominent topic, whereas “Differential Item Functioning” was the least well-known. The distribution of topics across academic journals emphasized the key role journals play in shaping the development and evolution of psychometric research. Through further exploration of topic correlations, potential future research directions and between-topic research areas were revealed. This study serves as a valuable resource for researchers aiming to keep up with the latest advancements in psychometrics. The findings provide crucial insights to guide and shape future research in the field.

## 1. INTRODUCTION

Psychometrics, although a field that has significantly advanced since the 2000s (Groenen & Ark, 2006), has much deeper historical roots. The foundations of psychometrics were established in the late 19th century by Sir Francis Galton, who aimed to evaluate human abilities using statistical methods and measurement techniques (Michell, 2022). In the post-World War II era, the emergence of psychometric methodologies and their applications in various domains contributed to noteworthy progress in the discipline (Jones & Thissen, 2006). In the mid-20th century, Charles Spearman introduced the concept of general intelligence, referred to as the “g

\*CONTACT: Kübra ATALAY KABASAKAL ✉ [katalay@hacettepe.edu.tr](mailto:katalay@hacettepe.edu.tr) 📍 Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

factor," and developed factor analysis, a fundamental method for identifying the common factors underlying psychological tests (Buckhalt, 1999).

This trajectory highlights how psychometric research in the 2000s gained momentum in response to the growing demands for psychological and educational measurement, paralleled by advancements in statistical and computational methods. During this period, particular emphasis was placed on the development and evaluation of psychometric tools, with researchers striving to ensure their reliability, validity, and applicability across diverse contexts (Martin & Savage-McGlynn, 2013). Simultaneously, interest in psychometric theory was revitalized, prompting academics to revisit foundational concepts and explore novel approaches to measurement and scale development (Jones & Thissen, 2006). This renewed focus drew attention to topics such as the psychometric validity of scales, item bias, differential item functioning, and estimation methods based on item response theory.

As researchers sought to bridge the gap between academic studies and practical applications, interest in the usability and interpretability of psychometric scales also increased (Vitoratou & Pickles, 2017). These advancements have enabled researchers to address increasingly complex and multifaceted research questions, thereby enhancing the depth and sophistication of psychometric analyses (Blanca *et al.*, 2018). By the 21st century, psychometric practices had become more sophisticated through the integration of technologies such as computer-assisted testing and data analytics. Psychometricians have utilized technological advancements to address emerging demands and develop the discipline. For example, the early 20th century witnessed a growing demand for standardized tests in education, marking a turning point in the development of measurement tools. During this period, pioneers like Thorndike emphasized the importance of measurement and evaluation practices in education and worked to establish a scientific foundation for these applications. The development and widespread adoption of educational achievement tests contributed significantly to the theoretical and practical growth of psychometrics.

Technological advancements, the growing demand for more sophisticated measurement tools, and an increasing emphasis on fairness and equity in psychological and educational assessment has shaped this progress in the field. For instance, innovations such as cognitive diagnostic modeling and adaptive testing have expanded the discipline's scope. However, the rise of large textual datasets has made organizing and understanding the themes in literature increasingly complex. Understanding the prominence of different themes in psychometric literature, how these themes have evolved over time, and how they vary across journals is crucial for evaluating the dynamic structure of the field. The interdisciplinary nature of psychometrics and its broad application in education, healthcare, and business further underscores the necessity of such an analysis. In this respect, topic modeling methods could be a desirable alternative for uncovering the hidden themes and trends in such a broad field.

Topic modeling, a powerful text mining technique that has become popular in natural language processing, can provide valuable insights into the themes and trends emerging in psychometrics literature (Gao & Sazara, 2023). This method's main goal is to identify underlying themes or topics within a large text corpus without prior content knowledge or labeling. One of the greatest advantages of topic modeling is its ability to process unstructured text data, which is ubiquitous in the digital age (Blei, 2012). The versatility of this technique makes it an essential tool for researchers and practitioners across a wide range of disciplines. Furthermore, its ability to handle diverse textual data, from short-form content like social media posts to long-form academic articles, underscores its adaptability (Richardson *et al.*, 2014). Boon-Itt and Skunkan (2020), for instance, leveraged topic modeling and sentiment analysis to understand public awareness of COVID-19 trends and identify significant themes of concern shared by Twitter users. Polatgil (2023) employed topic modeling to analyze user comments on the Duolingo mobile app, a widely utilized tool among language learners, with the aim of identifying the key aspects highlighted by users. Another study by Hwang *et al.* (2023) used topic modeling

techniques to identify research trends in published articles on the use of technology in mathematics education.

The application of topic modeling is increasingly gaining prominence in educational measurement. Anderson *et al.* (2020) introduced a novel approach to gather content-related validity evidence, incorporating topic modeling as a key method. Wheeler *et al.* (2024) stated that topic modeling is becoming more widespread in educational measurement research, particularly for analyzing responses to constructed-response items. A recent study by Xiong and Li (2023) employed topic modeling methods in the development of automatic scoring algorithms for constructed-response items. The growing use of topic modeling for various purposes in the literature, along with the need for a comprehensive examination of research trends in fields like psychometrics that involve large datasets, has led to the emergence of this study.

Despite the increasing use of topic modeling in educational and psychological research, there has been no comprehensive investigation of its application to psychometric literature. This presents a gap in understanding how core themes have evolved within the field and how they differ across publication venues. The lack of such analysis limits our ability to grasp the intellectual structure and thematic development of psychometrics as an interdisciplinary domain. Therefore, the present study is significant in that it systematically maps the landscape of psychometric research using structural topic modeling (STM), providing insights into its conceptual evolution and the distribution of topics across journals and time. Specifically, this study seeks to address the following questions:

- (1) What are the prominent themes in psychometric research?
- (2) How have these themes evolved over time?
- (3) How are these themes distributed across different journals?

In recent years, widespread adoption of technologies such as big data analytics and machine learning in psychometrics has significantly enhanced the field's capacity to perform complex analyses on large datasets. These technological advancements have also improved the validity and reliability of psychometric tools, enabling researchers to tackle increasingly complex challenges. Furthermore, the thematic organization and analysis of extensive literature have become crucial for guiding more focused and meaningful progress in the field. Such analyses not only provide valuable insights into the structure and focus of past research but also offer a framework for understanding future research orientations. To our knowledge, a comprehensive topic modeling and bibliometric analysis study in psychometrics remains absent. The only related example in the literature is the bibliometric investigation conducted by Zagaria and Lombardi (2024), which utilized the PsycINFO database to examine the relative prominence of Bayesian and frequentist approaches in the fields of psychology and psychometrics.

This study aims to explore the role of psychometrics in the scientific world and identify key focus areas for the future. It uses Structural Topic Modeling to analyze prominent themes in psychometric literature, their development over time, and their distribution across journals. This analysis provides a framework for understanding how fundamental psychometric methodologies have evolved and which themes have gained prominence in response to societal and technological changes. The thematic analysis further sheds light on the interdisciplinary nature of the field, highlighting the relationship between its theoretical foundations and practical applications. Together, these insights offer a more comprehensive understanding of the evolving landscape of psychometrics and its response to emerging demands and opportunities.

### 1.1. Structural Topic Modeling

Topic modeling is a machine learning approach that uses probabilistic models to uncover the themes and semantic patterns in large volumes of unstructured text data. By analyzing and linking documents based on word frequency patterns, these models can identify a set of



"topics," with each word and document associated with one or more topics (Blei, 2012). In topic modeling, each topic is defined by a collection of semantically related words. The model identifies these relationships by examining the frequency of words across the entire corpus of text. This enables a single word to relate to multiple topics, as its usage can vary depending on the context. Instead of assigning a single topic to each document, the model produces a probability distribution indicating the likelihood of a document belonging to each identified topic, based on the word patterns present (Blei *et al.*, 2003; Blei, 2012).

Structural topic modeling (STM) builds upon standard topic modeling by incorporating document metadata into the topic prediction process. This approach differs from traditional classification methods, which typically assign a document to a single, discrete category. STM is grounded in Latent Dirichlet Allocation (LDA; Blei *et al.*, 2003), but with a key distinction: LDA assumes topic prevalence and word usage patterns are static across all documents, while STM accounts for variability in these patterns by allowing them to be influenced by relevant covariates (Tonidandel *et al.*, 2021).

STM is a valuable tool for analyzing large volumes of unclassified text data. This advanced modeling approach identifies meaningful linguistic and semantic connections within the text and uncovers how these patterns vary across relevant metadata. STM's ability to link words with similar meanings and to distinguish multiple usages of the same word provides a more nuanced and contextual understanding of the content, making it particularly useful for analyzing extensive, unstructured datasets. By incorporating document metadata, STM enables researchers to examine the relationships between topic content, topic prevalence, and external variables (Tonidandel *et al.*, 2021). This technique can uncover hidden themes and trends in large text corpora and provide insights that may be missed by other methods.

Overall, STM is a powerful and versatile analytical framework for extracting meaningful insights from large, unstructured text datasets across a variety of contexts, including open-ended survey responses, news articles, and social media posts. This makes it an invaluable asset for researchers in fields such as social sciences, business, education, and public health (Bai *et al.*, 2021; Roberts *et al.*, 2014).

In the context of this study, STM offers a robust methodological framework for identifying the key themes within psychometric literature and examining how these themes evolve across time and publication venues. This approach is particularly valuable given the increasing volume and complexity of research in psychometrics, where traditional content analysis methods may fall short. By incorporating document-level metadata such as publication year and journal, STM enables a more nuanced exploration of thematic shifts in the field. Therefore, this study applies STM not only to map the conceptual landscape of psychometrics but also to uncover the dynamic interplay between methodological developments and thematic emphasis. Through this, we aim to contribute a systematic and scalable method for organizing knowledge in psychometrics and guiding future research directions.

## 2. METHOD

This study employed a descriptive and exploratory research design based on structural topic modeling (STM). The purpose was to examine the thematic structure of psychometric research by identifying latent topics within a large body of scholarly literature and exploring their temporal and journal-based variations. The following subsections outline the data collection, preprocessing, and analytical procedures used in the study.

### 2.1. Data Collection

We initiated our analysis of psychometric literature by identifying the most relevant journals in this field. Utilizing bibliometric techniques, we selected journals with a high volume of articles focused on psychometrics for further examination. To investigate the psychometrics publication landscape, we first identified the eleven most relevant journals in the field. After a

comprehensive review of the databases indexing these journals, we determined that Web of Science (WoS) was a primary source of bibliographic information. In the WoS search, we considered journals in both SSCI and ESCI indexes.

We retrieved the data on 26 January 2025 and extracted the publication name, publication year, journal name, document type, and abstract metadata from the WoS database and downloaded 1000 articles each time. We then examined the downloaded files to ensure they were from journal sources, the document type was an article, and an abstract was present in the metadata. After this examination, we consolidated the data into a single file. Our literature search identified 19,826 publications, but 10,829 of these lacked abstracts. This resulted in 8,997 publications remaining. An analysis of the publication years for the abstracted publications revealed that 46 were published before 1989, and these 46 were excluded from the corpus. Finally, the document type was restricted to articles, and this process left 8,523 articles. Information regarding 8,523 articles comprising the corpus is presented in Table 1. The names and abbreviations of the academic journals, the number of articles from each journal in the corpus, and the publication year ranges of the articles are presented in Table 1.

**Table 1.** *Features of the articles in the corpus.*

Journal Name and Abbrev	<i>n</i>	Publication Year Range
Applied Measurement in Education (AME)	549	1995- 2024
Applied Psychological Measurement (APM)	1009	1991- 2025
Educational and Psychological Measurement (EPM)	2146	1991- 2025
Educational Assessment (EA)	289	2005- 2024
Educational Measurement-Issues and Practice (EM-IP)	315	2013- 2024
International Journal of Assessment Tools in Education (IJATE)	395	2014- 2024
Journal of Educational and Behavioral Statistics (JEBS)	712	1994- 2025
Journal of Educational Measurement (JEM)	698	1992- 2025
Journal of Measurement and Evaluation in Education and Psychology (EPOD)	307	2010- 2024
Psychometrika (PSYCH)	1321	1990- 2024
Studies in Educational Evaluation (SEV)	782	2013- 2025

As shown in Table 1, most articles in the corpus are published in EPM, while the fewest originate from EA. Another noteworthy detail in Table 1 is that the EM-IP, IJATE, and EPOD journals, which show lower representations, are recent publications.

### 2.1.1. Data preparation

Before applying topic modeling, several text preprocessing steps were required. The text was converted to lowercase, and punctuation, numbers, and symbols were removed. Additionally, stop words (frequently used words) that provide little semantic meaning, such as "the", "a", "by", and "so", were eliminated. This is a widespread practice, as stop words appear frequently in text yet contribute little to the analysis. There is no universally accepted dictionary of stop words, but various libraries are available. For this application, the stop word lexicon from the tidytext R package was utilized (Silge & Robinson, 2016). We also removed terms like "approach", "data", "research", "article", "set", "procedure", and "framework" as sources of noise in the text, which led to the emergence of the Research Methods and Data Analysis topic. This text preprocessing step helped improve the performance of the language classification algorithm, such as STM, by focusing the analysis on more substantive and informative content within the text documents (Banks *et al.*, 2018).

Common text normalization techniques include stemming and lemmatization (Banks *et al.*, 2018). These methods reduce words to their roots but differ in their approaches. Stemming uses

pattern-based methods to determine the root without considering vocabulary, context, or parts of speech. In contrast, lemmatization involves morphological analysis to extract the root word while preserving semantic meaning (Singh & Gupta, 2017). Previous research has yielded mixed findings on the added benefits of applying stemming or lemmatization to large text datasets (Banks *et al.*, 2018). For this analysis, the lemmatization approach is selected as it maintains the conceptual significance of the words.

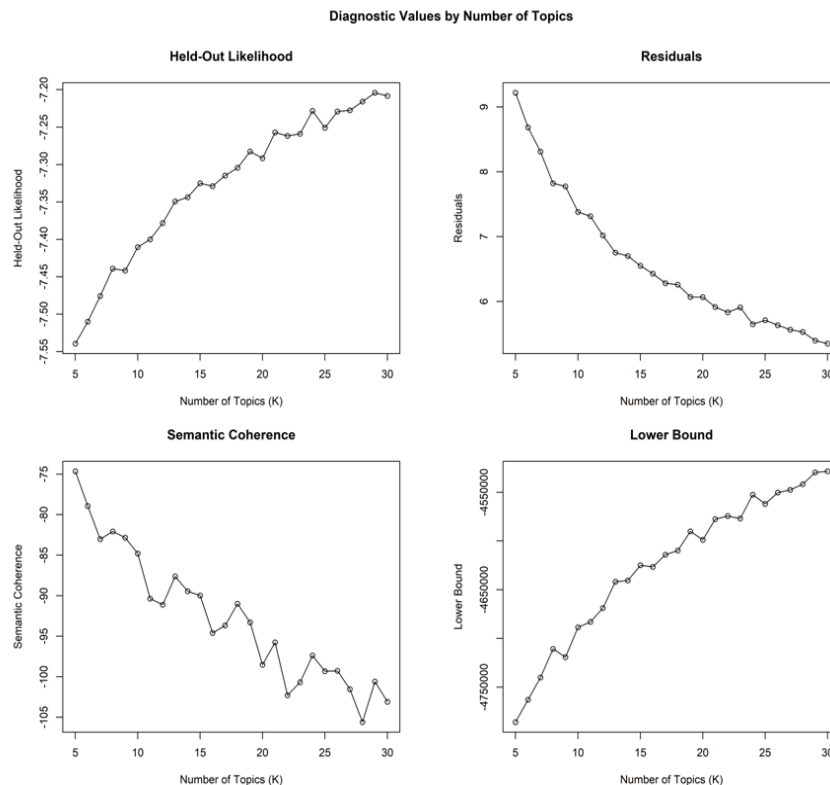
The article abstracts were pre-processed and converted into a traditional word-document matrix. In this matrix, each row represents the text from an individual abstract, and each column denotes a word used in the entire text corpus. In addition to single words, we also incorporated two-word combinations (bigrams) in our analyses. The decision to include unigrams, bigrams, and even trigrams reflects the need to balance adequately capturing meaning without unnecessarily increasing model complexity. As the n-gram size increases, the number of columns in the dataset grows exponentially, resulting in an extremely sparse, high-dimensional data structure. Unfortunately, high-dimensional data presents various analytical challenges that can impede the application of numerous techniques, a phenomenon known as the curse of dimensionality. While multiple n-grams should be avoided, they may be desirable if there is reason to believe they convey important semantic meaning beyond single words. In our case, we made an a priori decision to include bigrams because we believed that two-word combinations, such as “formative assessment,” “rater reliability,” or “internal consistency,” could reflect nuanced conceptual relationships. To mitigate the impact of sparse words, which contribute little to understanding common topics but can increase the computational complexity of structural topic models, we used a lower inclusion threshold. Only words or bigrams appearing in more than five documents were retained. Our final text corpus consisted of 8,523 documents, 13116 terms, and 384004 tokens.

### 2.1.2. Data analysis

The STM analysis was conducted using the *stm* package (Roberts *et al.*, 2019) in R software. The number of topics ( $K$ ) was determined by comparing models using semantic coherence and exclusivity metrics. Document metadata, including publication year and journal name, were incorporated as covariates to examine their influence on topic prevalence. The model output includes topic-word distributions ( $\beta$ ), document-topic distributions ( $\theta$ ), and estimates of topic variation over time and across journals.

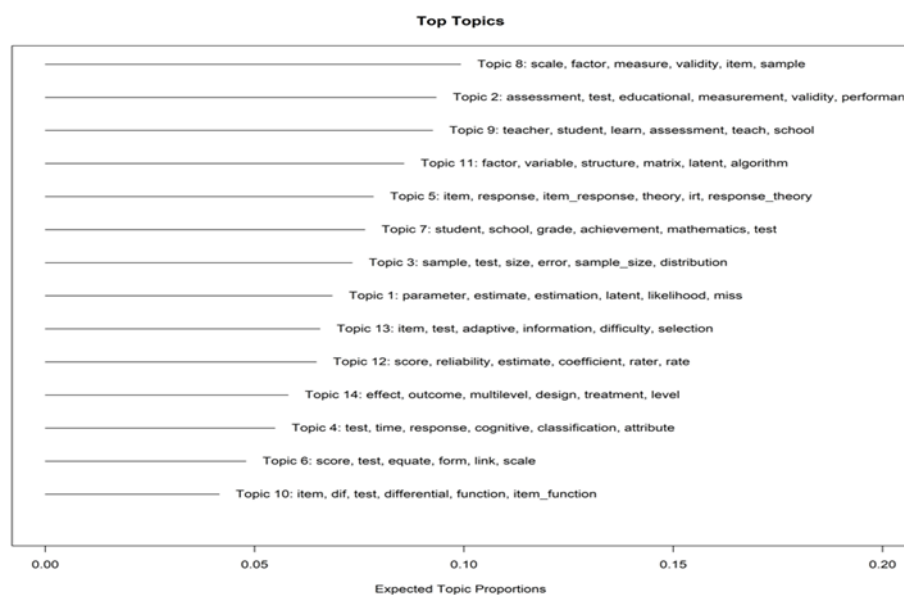
To determine the optimal number of topics, we used an iterative approach with the *searchK()* function in *stm*, estimating models with 5 to 30 topics. This process is analogous to inspecting a scree plot in exploratory factor analysis (Tonidandel *et al.*, 2021). As shown in Figure 1, key metrics guide topic selection. Semantic coherence measures how frequently a topic’s most probable terms co-occur, held-out likelihood assesses predictive performance on unseen data, residual variance indicates unexplained variation, and the lower bound reflects model log-likelihood, with higher values signifying better fit.

When determining the number of topics, the goal is to strike the best balance between coherence, exclusivity, and cohesion. As more topics are added, exclusivity tends to increase, enabling more refined differentiation. However, this also tends to reduce semantic coherence. The sweet spot is where coherence remains decent, but exclusivity is high. Based on our analysis, the models with 8 to 15 topics appeared to strike this balance. After reviewing these options, the 14-topic model was identified as the most meaningful and interpretable.

**Figure 1.** Comparison of solutions for topics spanning 5–30.

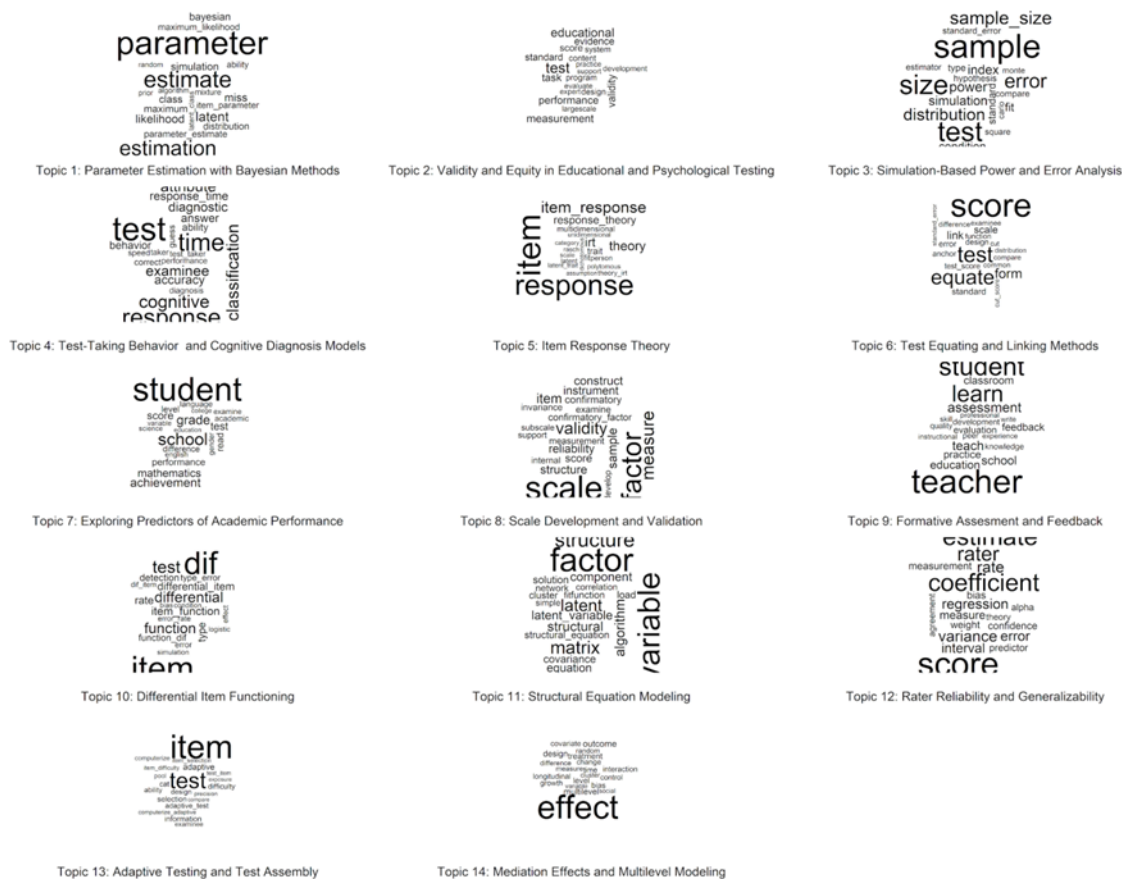
### 3. RESULTS

The analysis of 14 topics revealed their relative prevalence within the overall corpus, as depicted in Figure 2. This figure also presents the six most frequent words associated with each topic. Topic 8, characterized by the terms factor, scale, and measure, emerged as the most dominant topic. Conversely, Topic 10, containing the words item, differential, test, and function, demonstrated the lowest prevalence. The topic prevalence values span a range from 5% to 10%, with the four most dominant topics occupying the upper end of this spectrum, while the remaining topics have a prevalence of approximately 5% in the corpus, as can be seen in Figure 2.

**Figure 2.** Topics by prevalence.

STM uses metrics like prob (highest probability), FREX (frequency-exclusivity), lift, and score to uncover thematic patterns. Prob indicates the likelihood of a word belonging to a topic, while lift measures its uniqueness. FREX balances frequency and exclusivity, identifying words that are both common and distinctive (Roberts *et al.*, 2014). Score evaluates topic prevalence and distinctiveness, providing a comprehensive assessment of importance. In this study, topic labelling relied on high-probability words and FREX metrics, supplemented by expert review of the top five articles for each topic. Word clouds (Figure 3) visually represent the most probable words for each topic, with font size indicating probability.

**Figure 3.** Word clouds for each topic.



Word clouds illustrate the words with the highest probability of occurrence within each topic, where the font size corresponds to the probability of the word appearing. Figure 3 indicates that each topic is characterized by a unique set of related words. Furthermore, the most relevant words based on the FREX metric, along with the five most representative article abstracts for each topic, were examined in terms of context and depth of thematic content. The topics were then labeled based on the highest probability of word occurrence and the FREX metric. The topic naming process considered these metrics in conjunction with the content of the related articles, and the details are explained below.

The most prevalent topic identified in the analysis was Topic 8, Scale Development and Validation. This naming was derived from the five most representative article abstracts associated with this topic, which predominantly focus on the development, validation, and psychometric evaluation of various scales. For instance, Göral *et al.* (2024) conducted a methodological study to adapt and validate the Attitudes to Fertility and Childbearing Scale in a Turkish context, demonstrating strong reliability and validity through confirmatory factor analysis and internal consistency measures. Similarly, Tharenou and Terry (1998) assessed the reliability and validity of subjective and behavioral measures of managerial aspirations,



highlighting their distinct but related constructs and satisfactory psychometric properties. In another study, Tunç *et al.* (2021) developed the Hostility in Pandemic Scale to measure hostility levels during the COVID-19 pandemic, confirming its one-dimensional structure and high reliability through exploratory and confirmatory factor analyses. Lastly, Gregson (1991) explored the relationship between communication satisfaction and job satisfaction, using factor analysis to establish their separability as constructs. Collectively, these studies underscore the centrality of scale development and validation in psychometric research, as reflected in Topic 8.

The second most prevalent topic was Topic 2. Based on the FREX metrics and insights from articles in the corpus, this topic was titled as follows: Validity and Equity in Educational and Psychological Testing. This topic includes studies that examine the revision process and changes in the Standards for Educational and Psychological Testing, updated in 2014 (Plake and Wise, 2014), as well as research exploring how test results are evaluated as validity evidence (Cizek *et al.*, 2010). Additionally, discussions focus on how validity theory is shaped by the contexts in which tests are used (Sireci, 2013) and how core competencies in educational measurement can be developed (Ackerman, 2023). Furthermore, this topic emphasizes the need to consider assessment processes within a framework of social responsibility and justice (Buzick *et al.*, 2023) and discusses how the concept of validity can be expanded from a racial justice perspective (Lederman, 2023; Randall *et al.*, 2022). Studies in this topic also address the design and quality control processes of automated scoring systems (Rupp, 2018) and provide recommendations for the role of artificial intelligence in educational measurement, ensuring its use aligns with ethical standards (Briggs, 2024).

The third most prevalent topic was Topic 9 (Formative Assessment and Feedback). Brooks *et al.* (2020) evaluated the impact of a professional learning intervention using a student-centred feedback model in primary schools. Another study by Bastola and Hu (2021) examined students' views on feedback from thesis supervisors at a Nepalese university. Students felt the feedback was inadequate, but still engaged with it. The study also found that feedback engagement varied by discipline, suggesting the need for subject-specific feedback practices. Jiang and Ironsi (2024) explored how students respond to corrective peer feedback in the classroom. The results revealed that while students saw peer feedback as helpful, they also felt it was sometimes unfair or improperly assessed. These studies may demonstrate that the need for research on feedback in different fields and from various perspectives has led to this outcome.

The fourth most prevalent topic was Topic 5, Item Response Theory. This topic encompasses research focused on improving and refining IRT models, particularly concerning the analysis, scoring, and interpretation of data in educational testing and psychometrics. Studies such as those by Cohn & Huggins-Manley (2019), Huynh (1996), and Van Der Ark (2005) have contributed to advancements in these areas, emphasizing the development of more precise and reliable measurement techniques.

Topic 11, Structural Equation Modelling, represents another significant research focus on the corpus. Studies within this topic explore advanced methodologies in multivariate statistical analysis, particularly dimensionality reduction techniques, optimization algorithms, and methods for analysing complex, multi-dimensional data structures. Researchers such as Choi *et al.* (2016) and Kiers (1997) have contributed to the evolution of these techniques, aiming to develop more efficient ways to extract meaningful information from large datasets.

Topic 7, Exploring Predictors of Academic Performance, includes studies that investigate several factors influencing academic success. Some research explores the relationship between standardized test scores (e.g., SAT, GRE, GMAT) and academic performance in higher education (Meeter, 2022). On the other hand, some studies examine how socioeconomic status,

gender, and ethnicity affect student achievement at different educational levels, from primary school to university (Yavuz *et al.*, 2016).

Topic 3, Simulation-based Power and Error Analysis, includes studies that aim to improve and evaluate statistical methods used in hypothesis testing, multiple comparisons, and model evaluation. This research provides better tools and guidelines for conducting robust statistical analyses across different conditions and data types. For instance, MacDonald and Gardner (2000) used Monte Carlo methods to assess Type I error rates of six post hoc tests under various conditions. In a related approach, Guo and Luh (2008) proposed a method for determining appropriate sample sizes for Welch's F test in the presence of unequal variances through Monte Carlo simulations.

Topic 1, Parameter Estimation with Bayesian Methods, consists of studies discussing various methods for estimating item parameters in IRT models, including marginal Bayesian estimation, maximum likelihood estimation, and Gibbs sampling. These methodological advancements enhance the accuracy and applicability of IRT in psychometric research. In this context, Kim (2001, 2006) conducted studies on estimating item parameters in IRT models, particularly comparing calibration methods and evaluating the specific performance of MCMC-based Gibbs sampling for item and person parameter estimation.

Topic 13, Adaptive Testing and Test Assembly, focuses on advanced research in computerized adaptive testing (CAT). The studies explore methods for item selection, exposure control, and content balancing, aiming to enhance the efficiency, security, and fairness of adaptive tests while maintaining measurement precision (Chen & Lei, 2005; Han, 2012; Pan *et al.*, 2023).

Topic 12, Rater Reliability and Generalizability, encompasses research efforts to improve and evaluate various reliability and agreement measures used in psychometrics. Studies in this topic provide insights into the properties, limitations, and appropriate applications of these measures in different psychological and educational contexts. Some studies critique coefficient alpha for its limitations in handling correlated errors (Sijtsma & Pfadt, 2021; Rae, 2006), while another study discusses the relationships between weighted kappa, intraclass correlation, and product-moment correlation to better understand differences in interpretation (Schuster, 2004).

Topic 14, Mediation Effects and Multilevel Modeling, represents research in causal inference, focusing on methods for estimating causal effects in complex research designs (Park *et al.*, 2018; Talloen *et al.*, 2016). This topic addresses challenges such as confounding, noncompliance, and heterogeneity in treatment effects while providing tools for more robust causal inferences in fields such as education, social sciences, and medical research.

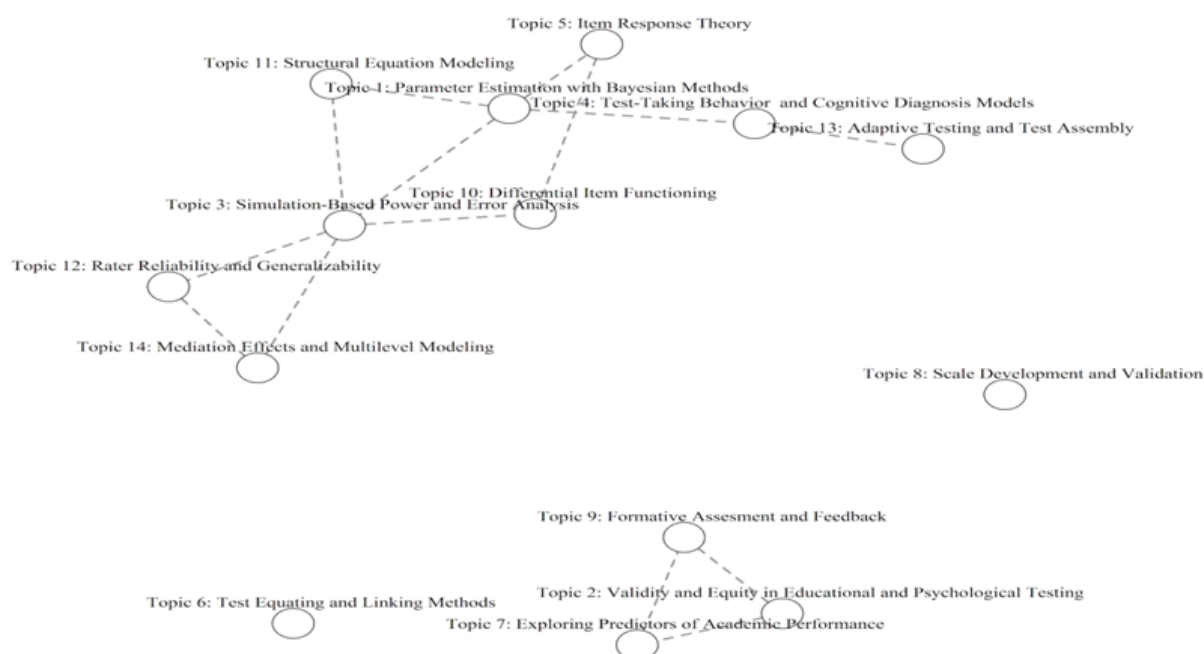
Topic 4, Test-Taking Behaviour and Cognitive Diagnosis Models, includes research on cognitive diagnostic assessment and modelling. Studies explore ways to improve the accuracy and interpretability of diagnostic information obtained from educational and psychological tests, tackling issues such as attribute specification, response time modelling, and integrating multiple data sources to enhance assessment precision. To give an example, a study by Zhan *et al.* (2022) proposes a multimodal joint cognitive diagnosis model that integrates accuracy, response times, and visual fixation counts from eye-tracking data. They also used an empirical example to demonstrate the applicability and benefits of the proposed model.

Topic 6, Test Equating and Linking Methods, focuses on comparing and refining test equating methods, understanding their theoretical foundations, and evaluating their performance under different conditions. This research aims to enhance the accuracy and reliability of test score comparisons across different forms and populations, ensuring fairer and more valid assessments. Numerous studies in the corpus examine the effectiveness of a wide range of equating techniques, often focusing on specific equating designs (Jiang *et al.*, 2012; Liu & Low, 2008; Von Davier *et al.*, 2004). Additionally, researchers also investigate the impact of sample size, population differences, and the type of anchor test used in the equating process (Liu & Low, 2008; Skaggs, 2005).

Topic 10, Differential Item Functioning (DIF), represents advanced research focused on improving the accuracy, power, and interpretability of DIF analyses in psychometric testing. The research aims to enhance the fairness and validity of educational and psychological assessments across distinct groups of test-takers. To this end, several studies have been conducted to assess the performance of common DIF detection techniques across different datasets and conditions and compare traditional methods with newer approaches (J. Chen *et al.*, 2013; Hidalgo & LÓpez-Pina, 2004; Shih & Wang, 2009).

The STM analysis enables the identification of correlated topics by analyzing their co-occurrence patterns within the same documents. As depicted in Figure 4, the topic network illustrates connections between nodes (topics) that exhibit a high probability of co-occurrence. Specifically, an edge is drawn between two nodes if their correlation coefficient exceeds 0.02. The observed topic correlations often reflect intuitive and meaningful relationships between the underlying concepts. The topic correlation analysis suggests that the psychometrics domain encompasses a distinct and diverse set of topics. For instance, Topic 8 (Scale Development and Validation) and Topic 6 (Test Equating and Linking Methods) were positioned separately from the other topics. Topic 2 (Validity and Equity in Educational and Psychological Testing), however, was linked to both Topic 7 (Exploring Predictors of Academic Performance) and Topic 9 (Formative Assessment and Feedback). Topic 1 (Parameter Estimation with Bayesian Methods) seemed related to several topics, including Topic 3 (Simulation-based Power and Error Analysis), Topic 4 (Test-taking Behavior and Cognitive Diagnosis Models), Topic 5 (Item Response Theory), and Topic 11 (Structural Equation Modeling). Topic 1 here can be regarded as a key concept that connects these topics due to its wide-ranging applications in psychometrics. Additionally, Topic 12 (Rater Reliability and Generalizability) and Topic 14 (Mediation Effects and Multilevel Modeling) were correlated, and they both were connected to Topic 3 (Simulation-based Power and Error Analysis). The graph also indicated that Topic 10 (Differential Item Functioning) shared connections with Topic 3 (Simulation-based Power and Error Analysis) and Topic 5 (Item Response Theory). This result is expected, as these three methods can be employed together to enhance test fairness and reliability.

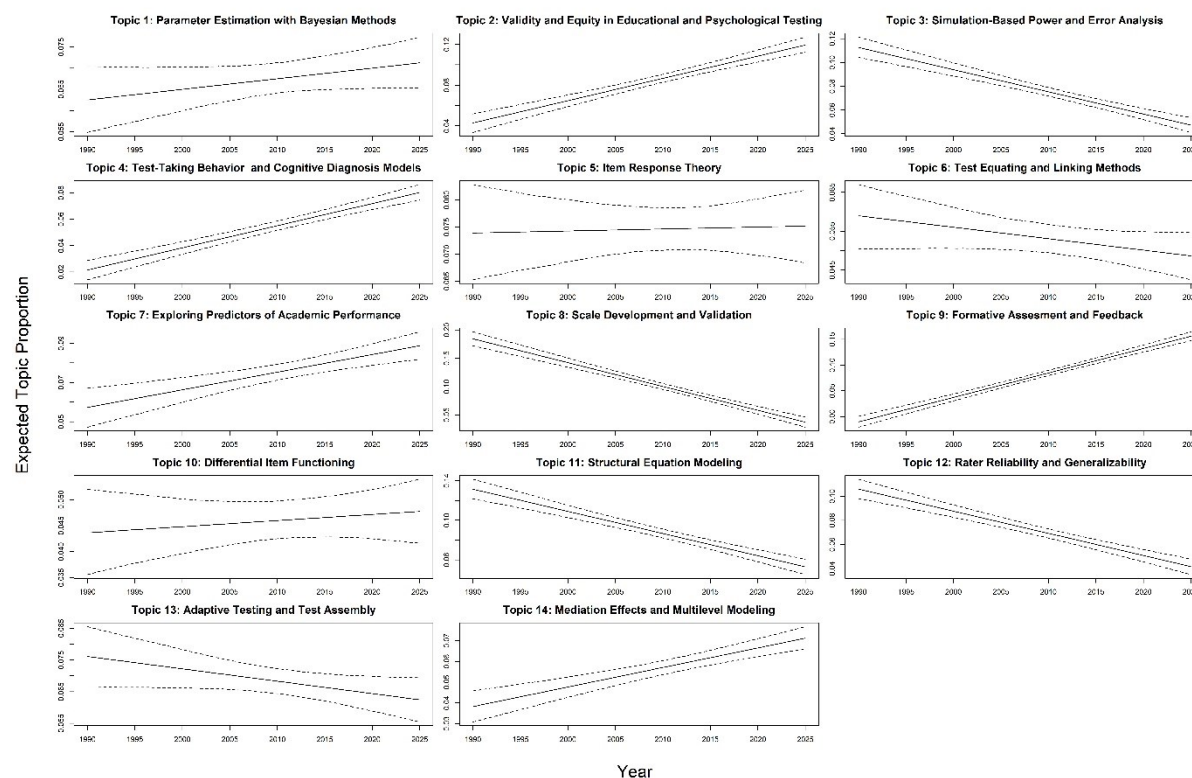
**Figure 4.** Network of topic correlation.



In the application of STM, the choice of covariates to include in the model is a crucial decision. In this study, the researchers selected the year of publication and journal names as covariates that may influence topic popularity. The key advantage of STM is its ability to investigate the

interactions between these covariates and the identified topics. By incorporating the publication year as a covariate, the model can track the prevalence of topics over time and enable comparative analysis. The topic prevalence is estimated as a function of the publication year, and confidence intervals around the estimated topic proportions are also generated. The temporal dynamics of topic popularity are presented in Figure 5, which illustrates the changes in the prevalence of each topic.

**Figure 5.** Expected topic proportions over time with %95 confidence intervals for each topic.



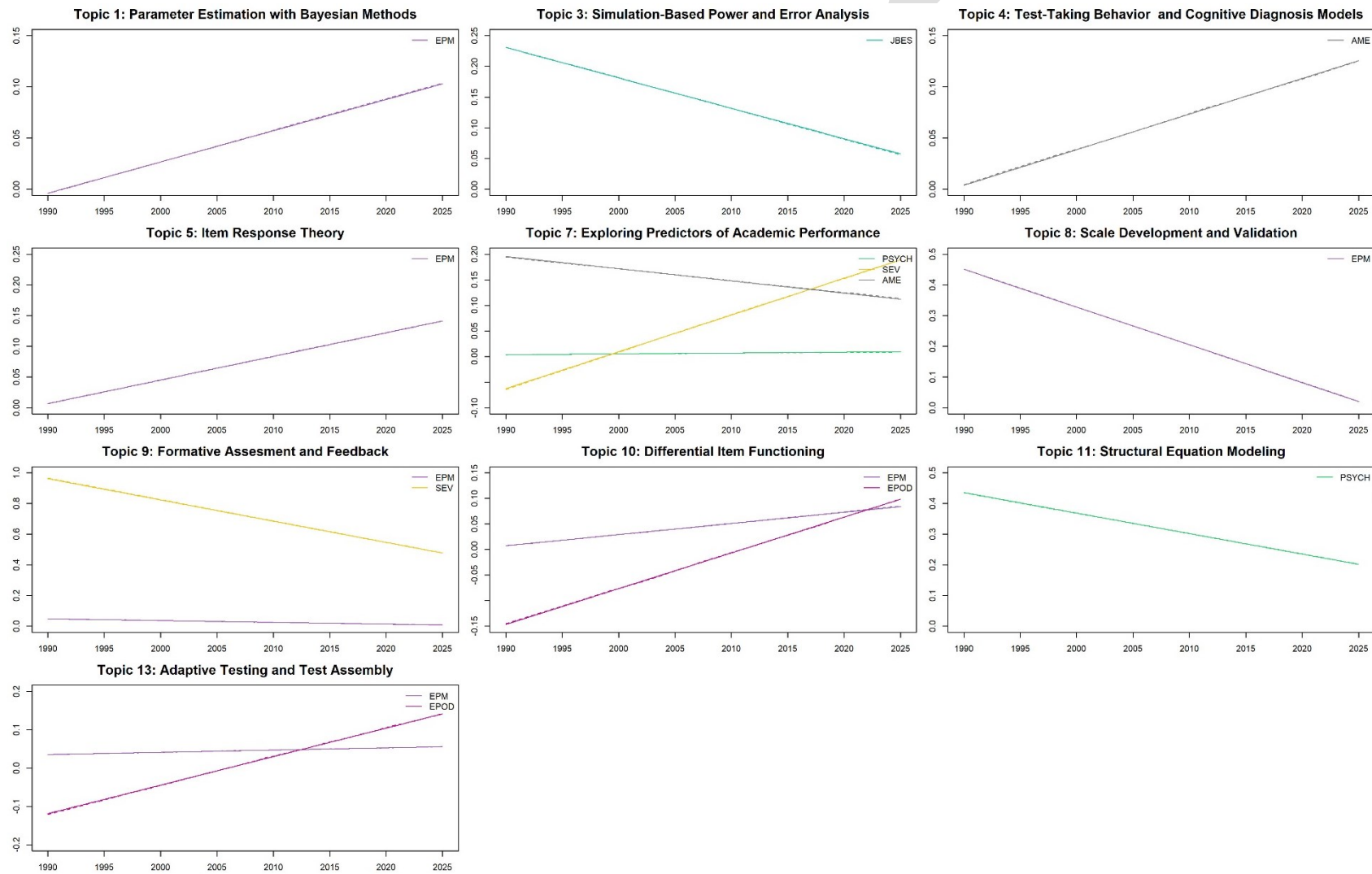
As shown in Figure 5, the researchers identified several "hot" and "cold" topics, reflecting those with increasing and decreasing prevalence trends, respectively, over the past 35 years. The "hot" topics include: Topic 2 (*Validity and Equity in Educational and Psychological Testing*), Topic 4 (*Test-Taking Behavior and Cognitive Diagnosis Models*), Topic 7 (*Exploring Predictors of Academic Performance*), Topic 9 (*Formative Assessment and Feedback*), and Topic 14 (*Mediation Effects and Multilevel Modeling*). These topics indicate growing interest and relevance in the field. In contrast, the "cold" topics include: Topic 3 (*Simulation-Based Power and Error Analysis*), Topic 8 (*Scale Development and Validation*), Topic 11 (*Structural Equation Modeling*), Topic 12 (*Rater Reliability and Generalizability*), and Topic 13 (*Adaptive Testing and Test Assembly*). These trends suggest a decline in research focus or interest in these areas over time. Additionally, several topics showed no significant relationship with the year of publication, indicating stable or inconsistent trends in their prevalence. These topics include: Topic 1 (*Parameter Estimation with Bayesian Methods*), Topic 5 (*Item Response Theory*), Topic 6 (*Test Equating and Linking Methods*), and Topic 10 (*Differential Item Functioning*). The lack of significant trends in these areas may reflect consistent but unchanging interest or methodological stability over the years.

After STM analysis, a regression model can be constructed with each document serving as the unit of analysis. The dependent variable is the proportion of each document associated with a particular topic in the STM model, while the independent variables are the document metadata. This approach enables examination of the main and interaction effects of the covariates after the STM analyses are conducted. The publication year variable considered a covariate in this study exhibited a significant effect on the topic proportions in all topics except for Topic 1

(*Parameter Estimation with Bayesian Methods*), Topic 5 (*Item Response Theory*), Topic 6 (*Test Equating and Linking Methods*), Topic 10 (*Differential Item Functioning*) and Topic 13 (*Adaptive Testing and Test Assembly*). Then, when the interaction effects of year and journal type were analyzed, it was observed that there were no significant interaction effects for Topic 2 (*Validity and Equity in Educational and Psychological Testing*), Topic 6 (*Test Equating and Linking Methods*) and Topic 12 (*Rater Reliability and Generalizability*) and Topic 14 (*Mediation Effects and Multilevel Modeling*). The graphs in Figure 6 illustrate the significant interaction effects between publication year and journal type. While not all journals are included for clarity, the analysis reveals several noteworthy trends. According to Figure 6, Topic 1 (*Parameter Estimation with Bayesian Methods*) has shown a significant increase over time in the EPM journal, although its overall prevalence remains low. Similarly, Topic 5 (*Item Response Theory*) has exhibited a rising trend in the EPM journal. In contrast, Topic 8 (*Scale Development and Validation*), despite being the most prevalent topic in the corpus, has experienced a notable decline in the EPM journal over the years.

Figure 6 also reveals that the JEBS journal has seen a pronounced decrease in the prevalence of Topic 3 (*Simulation-Based Power and Error Analysis*). While Topic 9 (*Formative Assessment and Feedback*) remains one of the most common topics, its prevalence has declined over time in the SEV journal, whereas it has remained stable in the EPM journal. On the other hand, Topic 10 (*Differential Item Functioning*), the least prevalent topic in the corpus, has shown a slight increase over time in the EPM journal. Topic 11 (*Structural Equation Modeling*), one of the most prevalent topics, has the highest representation in the EPM journal, but its prevalence in the PSYCH journal has gradually declined. Finally, Topic 13 (*Adaptive Testing and Test Assembly*), among the least prevalent topics, has demonstrated gradual growth in the EPOD and EPM journals.



**Figure 6.** Expected topic proportions of year and journal type interactions effect.

#### 4. DISCUSSION and CONCLUSION

This study conducted an in-depth analysis of the psychometric literature by using STM to identify thematic trends, the evolution of these dynamics over time, and the differences in topics covered across journals. To accomplish this, eleven leading journals in the field were included in the analysis. As a result, 14 topics were identified and labeled by the researchers. The findings revealed that Topic 8 (Scale Development and Validation) emerged as the most dominant topic, while Topic 10 (Differential Item Functioning) had the lowest prevalence in the corpus. The prevalence values of the topics ranged from 5% to 10%, with four topics occupying the upper end of this spectrum. These results highlight the major areas of focus in psychometric research and illustrate shifts in interest over time.

Previous applications of topic modeling in educational and psychological domains (e.g., Anderson *et al.*, 2020; Wheeler *et al.*, 2024; Xiong & Li, 2023) have focused on specific contexts such as constructed-response items, validity evidence, or automated scoring systems. However, to the best of our knowledge, no prior study has systematically applied STM to a comprehensive set of psychometric publications over time and across multiple journals. Compared to existing literature, the present study contributes a broader, field-wide perspective by combining bibliometric reach with the explanatory capacity of STM. This approach enables researchers to trace not only the dominant topics but also their temporal dynamics and journal-level distributions, thus offering a deeper understanding of psychometric scholarship.

Topic 8 (Scale Development and Validation) and Topic 2 (Validity and Equity in Educational and Psychological Testing) stand out by playing a vital role in both the theoretical and practical dimensions of psychometric research. This suggests that scale development and validation continue to be central to psychometric research, particularly given their fundamental role in ensuring robust measurement instruments. However, the interaction effect analysis revealed that while Topic 8 remains dominant, its prevalence has decreased over time in the EPM journal. This trend may indicate that some journals are shifting their focus towards newer methodologies and applications. On the other hand, emerging topics such as Topic 2 (Validity and Equity in Educational and Psychological Testing), Topic 4 (Test-Taking Behavior and Cognitive Diagnosis Models), and Topic 13 (Adaptive Testing and Test Assembly) underscore the increasing focus on technology-driven methodologies and individualized assessment practices. This shift aligns with broader advancements in fields such as educational technology and machine learning, where adaptive and personalized approaches are becoming increasingly important (van der Linden & Glas, 2000; Borsboom, 2005). However, the current study identifies Topic 9 (Formative Assessment and Feedback) and Topic 14 (Mediation Effects and Multilevel Modeling) as additional 'hot' topics, showing increasing prevalence over the years. This indicates a growing emphasis on assessment processes that prioritize formative evaluation and advanced statistical modeling techniques.

Topic 2, Validity and Equity in Educational and Psychological Testing, highlights the evolving nature of validity discourse in response to social, ethical, and technological advancements. While traditional validity frameworks primarily focused on test revisions and validation processes (Plake & Wise, 2014; Cizek *et al.*, 2010), recent discussions emphasize the broader societal implications of assessment practices (Buzick *et al.*, 2023). The integration of racial justice perspectives into validity theory (Lederman, 2023; Randall *et al.*, 2022) underscores the necessity of ensuring fairness and equity in educational and psychological measurement. Additionally, the increasing reliance on automated scoring systems (Rupp, 2018) and artificial intelligence-driven assessments (Briggs, 2024) calls for renewed scrutiny regarding the transparency, ethical use, and bias mitigation in these technologies. As validity continues to expand beyond psychometric properties to encompass social responsibility, future research should explore the intersection of validity theory with emerging AI methodologies, ethical assessment practices, and the broader implications of test fairness in diverse populations.

Despite the declining interest in "cold" topics such as Topic 8 (Scale Development and Validation), Topic 12 (Rater Reliability and Generalizability), and Topic 11 (Structural Equation Modeling), it is essential to explore how these areas can be better integrated into contemporary practices. These fundamental domains remain critical for the validity and reliability of measurement tools. Research could focus on revisiting these issues and updating them using modern technologies, particularly in high stakes testing contexts. Interestingly, while previous research suggested a consistent decline in simulation studies (Topic 3, Simulation-Based Power and Error Analysis), the present study shows that simulation remains a critical tool, albeit with reduced prominence compared to real-data-driven approaches. The decline in simulation studies is particularly notable in the JEBS journal, which may indicate a preference for empirical data over simulated conditions in recent years.

Another key finding of this study is the network of correlations between topics. Topic 2 (Validity and Equity in Educational and Psychological Testing) was found to be strongly linked to Topic 7 (Exploring Predictors of Academic Performance) and Topic 9 (Formative Assessment and Feedback), suggesting an interrelationship between test validity, student performance, and formative assessment practices. Furthermore, Topic 1 (Parameter Estimation with Bayesian Methods) was connected to multiple topics, including Topic 3 (Simulation-Based Power and Error Analysis), Topic 4 (Test-Taking Behavior and Cognitive Diagnosis Models), Topic 5 (Item Response Theory), and Topic 11 (Structural Equation Modeling), reinforcing its foundational role in psychometric methodologies. Topic 1 can be regarded as a key concept that connects these topics due to its wide-ranging applications in psychometrics. Additionally, Topic 12 (Rater Reliability and Generalizability) and Topic 14 (Mediation Effects and Multilevel Modeling) were correlated, and both were connected to Topic 3 (Simulation-Based Power and Error Analysis). Furthermore, Topic 10 (Differential Item Functioning) shared connections with Topic 3 (Simulation-Based Power and Error Analysis) and Topic 5 (Item Response Theory), which is an expected result given that these methods can be employed together to enhance test fairness and reliability.

The potential of artificial intelligence and big data technologies to enhance psychometric modeling and the development of measurement tools should be explored. For instance, integrating AI-based algorithms to understand more complex data structures, model response patterns, and improve prediction accuracy in large-scale testing is crucial. The ethical implementation, transparency, and reliability of these technologies should also be prioritized in scholarly discourse. These recommendations have the potential to facilitate the advancement of innovative research and effective practices within the field of psychometrics. By capitalizing on the opportunities presented by emerging technologies while maintaining a strong adherence to the theoretical foundations of the discipline, psychometrics could evolve into a more equitable, efficient, and accessible field.

To enhance the effective integration of psychometric findings into educational policies and practices, it is essential to establish stronger connections between academic research and applied studies. Guidance on the application of emerging methodologies, such as cognitive diagnostic modeling, is critical for promoting equity, fairness, and accessibility, especially in the context of educational assessment systems. This study has illustrated the interconnections between specific topics, such as item response theory and test equating and linking methods. Future research could focus on interdisciplinary projects that further strengthen these connections. For example, future research could examine how various psychometric methods can be simultaneously integrated within test development processes. Additionally, this study is confined to abstracts of articles published in eleven journals indexed in the WoS database. The selected journals are recognized as prominent within the fields of measurement, evaluation, and psychometrics. Replicating this study with a broader range of journals may yield different results.

The STM analysis also revealed significant interaction effects between publication year and journal type. For instance, Topic 1 (Parameter Estimation with Bayesian Methods) and Topic 5 (Item Response Theory) have shown significant increases over time, particularly in the EPM journal. In contrast, Topic 11 (Structural Equation Modeling), despite its initial prominence, has exhibited a decline in the PSYCH journal. These findings suggest that the prevalence of specific psychometric topics may vary across journals, reflecting different editorial priorities and emerging research trends.

The widespread adoption of technologies such as big data analytics and machine learning has enhanced the accessibility and meaningfulness of real data analysis. Despite the decline in simulation studies, their critical role in strengthening theoretical and methodological foundations should not be overlooked. It is clear that simulation continues to be a critical tool, particularly for the testing and validation of new methodologies. In the future, establishing a balance between simulation studies and real data analyses may provide an approach that fosters both theoretical rigor and practical utility in psychometric research. In this regard, the growing prominence of real data could encourage the integration of simulation studies with more realistic scenarios and hybrid methods, thereby contributing to the development of more robust and comprehensive analytical frameworks within the field of psychometrics.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

### Contribution of Authors

**Kübra Atalay Kabasakal:** Research Design, Resources, Methodology, Visualization, Software, Analysis, and Writing-original draft. **Duygu Koçak:** Research Design, Systematic review, Methodology, Supervision, and Critical Review. **Rabia Akcan:** Research Design, Systematic review, Methodology, Supervision, and Critical Review.

### Orcid

Kübra Atalay Kabasakal  <https://orcid.org/0000-0002-3580-5568>

Duygu Koçak  <https://orcid.org/0000-0003-3211-0426>

Rabia Akcan  <https://orcid.org/0000-0003-3025-774X>

### REFERENCES

- Ackerman, T.A., Bandalos, D.L., Briggs, D.C., Everson, H.T., Ho, A.D., Lottridge, S.M., Madison, M.J., Sinharay, S., Rodriguez, M.C., Russell, M., Von Davier, A.A., & Wind, S.A. (2023). Foundational competencies in educational measurement. *Educational Measurement Issues and Practice*, 43(3), 7–17. <https://doi.org/10.1111/emip.12581>
- Anderson, D., Rowley, B., Stegenga, S., Irvin, P.S., & Rosenberg, J.M. (2020). Evaluating content-related validity evidence using a text-based machine learning procedure. *Educational Measurement: Issues and Practice*, 39(4), 53–64. <https://doi.org/10.1111/emip.12314>
- Bai, X., Zhang, X., Li, K.X., Zhou, Y., & Yuen, K.F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, 102, 11–24. <https://doi.org/10.1016/j.tranpol.2020.12.013>
- Banks, G.C., Woznyj, H.M., Wesslen, R.S., & Ross, R.L. (2018). A review of best practice recommendations for text analysis in R (and a User-Friendly app). *Journal of Business and Psychology*, 33(4), 445–459. <https://doi.org/10.1007/s10869-017-9528-3>
- Bastola, M. N., & Hu, G. (2021). Chasing my supervisor all day long like a hungry child seeking her mother!: Students' perceptions of supervisory feedback. *Studies in Educational Evaluation*, 70, 101055. <https://doi.org/10.1016/j.stueduc.2021.101055>

- Blanca, M.J., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2018). Non-normal data: Is ANOVA still a valid option? *Psicothema*, 30(4), 552-557. <https://doi.org/10.7334/psicothema2018.245>
- Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://doi.org/10.5555/944919.944937>
- Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>
- Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4), e21978.
- Briggs, D.C. (2024). Strive for measurement, set new standards, and try not to be evil. *Journal of Educational and Behavioral Statistics*, 49(5), 694-701. <https://doi.org/10.3102/10769986241238479>
- Brooks, C., Burton, R., Van Der Kleij, F., Carroll, A., Olave, K., & Hattie, J. (2020). From fixing the work to improving the learner: An initial evaluation of a professional learning intervention using a new student-centred feedback model. *Studies in Educational Evaluation*, 68, 100943. <https://doi.org/10.1016/j.stueduc.2020.100943>
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511613980>
- Buckhalt, J.A. (1999). Defending the science of mental ability and its central dogma. Review of Jensen on Intelligence-g-Factor. *Psychology*, 10(23). <http://www.cogsci.ecs.soton.ac.uk/cgi/psyc/newpsy?10.47>
- Buzick, H.M., Casabianca, J.M., & Gholson, M.L. (2023). Personalizing Large-Scale Assessment in practice. *Educational Measurement Issues and Practice*, 42(2), 5–11. <https://doi.org/10.1111/emip.12551>
- Chen, S., & Lei, P. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, 29(3), 204-217. <https://doi.org/10.1177/0146621604271495>
- Chen, J., Chen, C., & Shih, C. (2013). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, 38(1), 18-36. <https://doi.org/10.1177/0146621613488643>
- Choi, J.Y., Hwang, H., Yamamoto, M., Jung, K., & Woodward, T.S. (2016). A unified approach to functional principal component analysis and functional Multiple-Set canonical correlation. *Psychometrika*, 82(2), 427–441. <https://doi.org/10.1007/s11336-015-9478-5>
- Cizek, G.J., Bowen, D., & Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: a Follow-Up Study. *Educational and Psychological Measurement*, 70(5), 732–743. <https://doi.org/10.1177/0013164410379323>
- Cohn, S., & Huggins-Manley, A.C. (2019). Applying unidimensional models for semiordeed data to scale data with neutral responses. *Educational and Psychological Measurement*, 80(2), 242–261. <https://doi.org/10.1177/0013164419861143>
- Jones, L.V., & Thissen, D.M. (2006). *A history and overview of psychometrics*. In Handbook of statistics (pp. 1–27). [https://doi.org/10.1016/s0169-7161\(06\)26001-2](https://doi.org/10.1016/s0169-7161(06)26001-2)
- Gao, X., & Sazara, C. (2023). Discovering mental health research topics with topic modeling. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2308.13569>
- Göral, S., Özkan, S., Sercekus, P., & Alataş, E. (2021). The validity and reliability of the Turkish version of the Attitudes to Fer-Tility and Childbearing Scale (AFCS). *International Journal of Assessment Tools in Education*, 8(4), 764-774. <https://doi.org/10.21449/ijate.773132>
- Gregson, T. (1991). The separate constructs of communication satisfaction and job satisfaction. *Educational and Psychological Measurement*, 51(1), 39-48. <https://doi.org/10.1177/0013164491511003>



- Groenen, P.J.F., & van der Ark, L.A. (2006). Visions of 70 years of psychometrics: the past, present, and future. *Statistica Neerlandica*, 60(2), 135–144. <https://doi.org/10.1111/j.1467-9574.2006.00318.x>
- Guo, J., & Luh, W. (2008). Approximate sample size formulas for testing group mean differences when variances are unequal in One-Way ANOVA. *Educational and Psychological Measurement*, 68(6), 959–971. <https://doi.org/10.1177/0013164408318759>
- Hidalgo, M.D., & LÓpez-Pina, J.A. (2004). Differential Item Functioning Detection and Effect Size: A Comparison between Logistic Regression and Mantel-Haenszel Procedures. *Educational and Psychological Measurement*, 64(6), 903-915. <https://doi.org/10.1177/0013164403261769>
- Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika*, 61(1), 31-39. <https://doi.org/10.1007/bf02296957>
- Hwang, S., Flavin, E., & Lee, J.E. (2023). Exploring research trends of technology use in mathematics education: A scoping review using topic modeling. *Education and Information Technologies*, 28, 10753–10780. <https://doi.org/10.1007/s10639-023-11603-0>
- Jiang, Y., Von Davier, A.A., & Chen, H. (2012). Evaluating equating results: percent relative error for chained kernel equating. *Journal of Educational Measurement*, 49(1), 39–58. <https://doi.org/10.1111/j.1745-3984.2011.00159.x>
- Jiang, X., & Ironsi, S.S. (2024). Do learners learn from corrective peer feedback? Insights from students. *Studies in Educational Evaluation*, 83, 101385. <https://doi.org/10.1016/j.stueduc.2024.101385>
- Kiers, H.A.L. (1997). Three-mode orthomax rotation. *Psychometrika*, 62(4), 579–598. <https://doi.org/10.1007/bf02294644>
- Kim, S. (2001). An evaluation of a Markov Chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, 25(2), 163-176. <https://doi.org/10.1177/01466210122031984>
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43(4), 355–381. <https://doi.org/10.1111/j.1745-3984.2006.00021.x>
- Lederman, J. (2023). Validity and racial justice in educational assessment. *Applied Measurement in Education*, 36(3), 242-254. <https://doi.org/10.1080/08957347.2023.2214654>
- Liu, J., & Low, A.C. (2008). A Comparison of the Kernel Equating Method with Traditional Equating Methods Using SAT® Data. *Journal of Educational Measurement*, 45(4), 309–323. <https://doi.org/10.1111/j.1745-3984.2008.00067.x>
- MacDonald, P.L., & Gardner, R.C. (2000). Type I Error Rate Comparisons of Post Hoc Procedures for I j Chi-Square Tables. *Educational and Psychological Measurement*, 60(5), 735–754. <https://doi.org/10.1177/00131640021970871>
- Martin, C.R., & Savage-McGlynn, E. (2013). A ‘good practice’ guide for the reporting of design and analysis for psychometric evaluation. *Journal of Reproductive and Infant Psychology*, 31(5), 449–455. <https://doi.org/10.1080/02646838.2013.835036>
- Meeter, M. (2022). Predicting Retention in Higher Education from high-stakes Exams or School GPA. *Educational Assessment*, 28(1), 1-10. <https://doi.org/10.1080/10627197.2022.2130748>
- Michell, J. (2022). The art of imposing measurement upon the mind: Sir Francis Galton and the genesis of the psychometric paradigm. *Theory & Psychology*, 32(3), 375-400. <https://doi.org/10.1177/09593543211017671>
- Pan, Y., Livne, O., Wollack, J.A., & Sinharay, S. (2023). Item selection algorithm based on collaborative filtering for item exposure control. *Educational Measurement Issues and Practice*, 42(4), 6–18. <https://doi.org/10.1111/emip.12578>

- Park, S., Steiner, P.M., & Kaplan, D. (2018). Identification and sensitivity analysis for average causal mediation effects with time-varying treatments and mediators: Investigating the underlying mechanisms of kindergarten retention policy. *Psychometrika*, 83(2), 298–320. <https://doi.org/10.1007/s11336-018-9606-0>
- Plake, B.S., & Wise, L.L. (2014). What is the role and importance of the revised AERA, APA, NCME standards for educational and psychological testing? *Educational Measurement Issues and Practice*, 33(4), 4–12. <https://doi.org/10.1111/emip.12045>
- Polatgil, M. (2023). Analyzing comments made to the Duolingo mobile application with topic modeling. *International Journal of Computing and Digital Systems*, 13(1), 223–230.
- Randall, J., Slomp, D., Poe, M., & Oliveri, M.E. (2022). Disrupting White Supremacy in Assessment: Toward a Justice-Oriented, Antiracist validity framework. *Educational Assessment*, 27(2), 170–178. <https://doi.org/10.1080/10627197.2022.2042682>
- Richardson, G.M., Bowers, J., Woodill, A.J., Barr, J.R., Gawron, J.M., & Levine, R.A. (2014). Topic models: A tutorial with R. *International Journal of Semantic Computing*, 8(01), 85–98.
- Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., & Rand, D.G. (2014). Structural topic models for Open-Ended Survey Responses. *American Journal of Political Science*, 58(4), 1064–1082. <https://doi.org/10.1111/ajps.12103>
- Roberts, M.E., Stewart, B.M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2). <https://doi.org/10.18637/jss.v091.i02>
- Rupp, A.A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, 31(3), 191–214. <https://doi.org/10.1080/08957347.2018.1464448>
- Schuster, C. (2004). A Note on the Interpretation of Weighted Kappa and its Relations to Other Rater Agreement Statistics for Metric Scales. *Educational and Psychological Measurement*, 64(2), 243–253. <https://doi.org/10.1177/0013164403260197>
- Shih, C., & Wang, W. (2009). Differential Item Functioning Detection Using the Multiple Indicators, Multiple Causes Method with a Pure Short Anchor. *Applied Psychological Measurement*, 33(3), 184–199. <https://doi.org/10.1177/0146621608321758>
- Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open-Source Software*, 1(3), 37. <https://doi.org/10.21105/joss.00037>
- Singh, J., & Gupta, V. (2017). A systematic review of text stemming techniques. *Artificial Intelligence Review*, 48(2), 157–217. <https://doi.org/10.1007/s10462-016-9498-2>
- Sireci, S.G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50(1), 99–104. <https://doi.org/10.1111/jedm.12005>
- Sijtsma, K., & Pfadt, J.M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, 86(4), 843–860.
- Tharenou, P., & Terry, D.J. (1998). Reliability and validity of scores on scales to measure managerial aspirations. *Educational and Psychological Measurement*, 58(3), 475–492. <https://doi.org/10.1177/0013164498058003008>
- Talloe, W., Moerkerke, B., Loeys, T., De Naeghel, J., Van Keer, H., & Vansteelandt, S. (2016). Estimation of indirect effects in the presence of unmeasured confounding for the Mediator–Outcome relationship in a multilevel 2-1-1 mediation model. *Journal of Educational and Behavioral Statistics*, 41(4), 359–391. <https://doi.org/10.3102/1076998616636855>
- Tonidandel, S., Summerville, K.M., Gentry, W.A., & Young, S.F. (2021). Using structural topic modeling to gain insight into challenges faced by leaders. *The Leadership Quarterly*, 33(5), 101576. <https://doi.org/10.1016/j.leaqua.2021.101576>

- Tunç, E.B., Parlak, S., Uluman, M., & Eryiğit, D. (2021). Development of the Hostility in Pandemic Scale (HPS): A Validity and Reliability study. *International Journal of Assessment Tools in Education*, 8(3), 475–486. <https://doi.org/10.21449/ijate.837616>
- Wheeler, J.M., Cohen, A.S., & Wang, S. (2024). A comparison of latent semantic analysis and latent Dirichlet allocation in educational measurement. *Journal of Educational and Behavioral Statistics*, 49(5), 848–874. <https://doi.org/10.3102/10769986231209446>
- Van Der Ark, L.A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika*, 70(2), 283–304. <https://doi.org/10.1007/s11336-000-0862-3>
- Van der Linden, W.J., & Glas, C.A.W. (2000). *Computerized adaptive testing: Theory and practice*. Springer. <https://doi.org/10.1007/978-1-4757-3224-0>
- Vitoratou, S., & Pickles, A. (2017). Psychometric analysis of the Mental Health Continuum-Short Form. *Journal of Clinical Psychology*, 73(10), 1307-1322. <https://doi.org/10.1002/jclp.22422>
- Xiong, J., & Li, F. (2023). Bilevel topic model-based multitask learning for constructed-response multidimensional automated scoring and interpretation. *Educational Measurement: Issues and Practice*, 42(2), 42–61. <https://doi.org/10.1111/emip.12550>
- Yavuz, S., Odabaş, M., & Özdemir, A. (2016). Öğrencilerin sosyoekonomik düzeylerinin TEOG matematik başarısına etkisi [Effect of socio-economic status on student's TEOG mathematics achievement]. *Journal of Measurement and Evaluation in Education and Psychology*, 7(1), 85–95. <https://doi.org/10.21031/epod.86531>
- Zagaria, A., & Lombardi, L. (2024). Bayesian versus frequentist approaches in psychometrics: a bibliometric analysis. *Discover Psychology*, 4, 61. <https://doi.org/10.1007/s44202-024-00164-z>
- Zhan, P., Man, K., Wind, S.A., & Malone, J. (2022). Cognitive diagnosis modeling incorporating response times and fixation counts providing comprehensive feedback and accurate diagnosis. *Journal of Educational and Behavioral Statistics*, 47(6), 736–776. <https://doi.org/10.3102/10769986221111085>

## Support for Gender Equality among Men Scale: Adaptation to Turkish culture

Esma Esen Çiftçi<sup>1\*</sup>, Esra Daşçı<sup>2</sup>, Cansu Ayan<sup>3</sup>, Zeynep Uludağ<sup>4</sup>

<sup>1</sup>Anadolu University, Faculty of Humanities, Department of Psychology, Eskişehir, Türkiye

<sup>2</sup>Forward College, Department of Psychology, Lisbon, Portugal

<sup>3</sup>Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Ankara, Türkiye

<sup>4</sup>Ardahan University, Faculty of Humanities and Letters, Department of Psychology, Ardahan, Türkiye

### ARTICLE HISTORY

Received: Mar. 22, 2025

Accepted: July 8, 2025

### Keywords:

Gender equality,  
Participation of men in  
gender equality,  
Adaptation,  
Support for gender  
equality scale.

**Abstract:** Men's participation is an important indicator for achieving gender equality. The purpose of this study is to adapt the “Support for Gender Equality among Men Scale (SGEMS)” developed by Sudkamper *et al.* (2020) to Turkish. The scale examines men's support for gender equality in two sub-dimensions: public space and household. In the study, 419 cis-heterosexual men participated. The confirmatory factor analysis results, which were used for validity evidence, showed that the Turkish version of SGEMS had the same two-factor structure as the original form. To conduct criterion-based validity, the participants answered the Ambivalent Sexism Inventory (ASI) and the Gender Equality in Turkish Men Scale (GEMS) along with the SGEMS. A significant negative correlation was found between the total score obtained from the SGEMS and the total score obtained from ASI, and a significant positive correlation between the total score obtained from SGEMS and the GEMS. While Cronbach's  $\alpha$  internal consistency reliability coefficient for the entire SGEMS was .89, the internal consistency coefficients for the Public and Household subscales were .89 and .78, respectively. Finally, it was examined whether the scale showed measurement invariance for two different age groups, and it was found that configural metric and scalar invariance were achieved. In conclusion, the SGEMS has been introduced to the literature as a valid and reliable scale measuring men's support for gender equality in the Turkish sample.

## 1. INTRODUCTION

Despite numerous efforts to achieve gender equality, significant barriers remain, particularly regarding women's participation in the labor force and career advancement (Bear *et al.*, 2025; Eagly & Karau, 2002; Heilman, 2012), perpetuating economic inequalities. Additionally, women bear a disproportionate burden in domestic work (Cho *et al.*, 2025). Historically, gender inequality research has focused on women's empowerment as the solution (Belingheri *et al.*, 2021; Ryan & Kirby, 2018). While women's efforts are essential, achieving meaningful change requires the inclusion of men, who are key agents in perpetuating inequalities. Recent gender

\*CONTACT: Esma Esen ÇİFTÇİ ✉ [e.ciftci@anadolu.edu.tr](mailto:e.ciftci@anadolu.edu.tr) 📍 Anadolu University, Faculty of Humanities, Department of Psychology, Eskişehir, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

studies emphasize the importance of men's participation in gender equality for meaningful social change (Moser & Branscombe, 2022; van Laar *et al.*, 2024). Therefore, excluding men from these efforts is no longer viable.

Indeed, research suggests that instruments that measure women's attitudes and behaviors towards gender equality are insufficient to measure men's participation in gender equality because both men and women have different ways of supporting gender equality and the reactions they receive from others when they support gender equality are different (Anderson, 2009; Cihangir *et al.*, 2014; Drury & Kaiser, 2014). For this reason, Sudkamper *et al.* (2020) developed the Support for Gender Equality among Men Scale (SGEMS). The current study will examine the adaptation of this scale to Turkish.

### 1.1. Related Literature and Existing Scales on Men's Participation in Gender Equality

Participation in gender equality could be in the attitudinal dimension as well as in the behavioral dimension. While the attitudinal dimension may involve opposing sexist attitudes, the behavioral dimension may involve the tendency to actively exhibit behaviors that will eliminate sexism (Manstead & Parker, 1995). Most frequently used scales in the literature, such as the Liberal Feminist Attitudes and Ideology Scale (Morgan, 1996), Traditional-Egalitarian Gender Roles Scale (Larsen & Long, 1988), Attitudes towards Women Scale (Spence *et al.*, 1973), Gender Role Stereotypes Scale (Mills *et al.*, 2012), 'Gender Role Beliefs Scale' (Brown & Gladstone, 2012), typically reveals instruments that assess the attitudinal aspect of gender equality. In addition, the discussion regarding the measurement tools for gender equality is limited to the answers to the question 'how should gender equality be?'. However, there was no questioning about how men behave in situations of inequality.

Measuring only attitudes toward gender equality may be insufficient to fully assess participation in it, and there is a lack of scales focusing on men's behavioral actions to promote equality beyond attitudes (Ajzen, 1991; Branscombe & Deaux, 1991; Woodzicka & LaFrance, 2001; Zucker, 2004). Therefore, there is a need to measure behavioral steps taken by men to ensure gender equality rather than attitudes developed against sexist ideologies or stereotypes (Ajzen & Sheikh, 2013; Brown & Gladstone, 2012).

Some studies do not aim to measure men's participation in gender equality, including behavioral measurements that will also serve this purpose. For example, Cihangir *et al.* (2014) measured whether male employees react on behalf of women exposed to gender inequality in the workplace. Similarly, Iyer and Ryan (2009) measured how much men support positive discrimination to ensure gender equality in the workplace. However, as mentioned earlier, there is still a need to develop a scale that measures behavioral disposition towards gender equality, focusing on men.

In addition, most studies have only analyzed men's contribution to gender equality in the workplace. However, four main points become evident in men's attitudes and behaviors towards achieving gender equality in public spaces: Men participate in political activities for gender equality (Iyer & Ryan, 2009; White, 2006); men object when they see gender inequality (Cihangir *et al.*, 2014; Eliezer & Major, 2011; Rasinski & Czopp, 2010); men develop discourse on gender equality (Kaufman & Kimmel, 2011; Lemaster *et al.*, 2015); men support a more inclusive organizational culture (Liff & Cameron, 1997). These findings demonstrate that men's efforts toward gender equality are not confined to professional settings but manifest in broader public and social contexts.

Moreover, studies that measure men's gender equality generally focus on gender equality in the workplace and mostly overlook the domestic factor (Iyer & Ryan, 2009). Within the household, men play a crucial role in promoting gender equality by treating their partners with respect (Frei & Shaver, 2002; Hendrick & Hendrick, 2006), sharing household chores equally (Deutsch, 1999; Dotti Sani, 2014; Kosakowska-Brezecka *et al.*, 2016), and sharing parenting and childcare tasks (Deutsch, 1999; Gartner, 2007; Haas, 2003). However, the existing studies often



require participants to document their engagement in household tasks through diary records (Achen & Stafford, 2005; Bianchi *et al.*, 2000; Craig *et al.*, 2016) or respond to direct questions regarding their participation in domestic responsibilities, such as the frequency of changing their child's diaper (Kato-Wallece *et al.*, 2014). Merely participating in household chores and childcare does not fully achieve gender equality at home. A truly equitable household requires a shift in power dynamics, shared decision-making, and a deeper commitment to dismantling ingrained gender norms that shape domestic responsibilities.

Given all this, SGEMS (Sudkamper *et al.*, 2020) is a psychometric instrument designed to assess men's behavioral intentions and attitudes toward gender equality, specifically in both public and domestic spheres. The scale measures men's participation in promoting gender equality through two primary sub-dimensions: involvement in gender equality in public spaces and within the household. It aims to capture men's support for gender equality behaviours and their alignment with gender-equal practices, distinguishing it from scales focusing solely on attitudes toward women or sexism.

## 1.2. Existing Scales on Men's Participation in Gender Equality in Turkey

While international literature explores men's participation in gender equality, social psychology research in Turkey lacks sufficient focus on this topic, with related studies primarily addressing violence against women and men's attitudes toward gender roles.

The scales of violence against women are Attitudes towards Violence against Women in Marriage Scale (Sakallı-Uğurlu & Ulu, 2003), Intimate Violence Responsibility Scale (Akin *et al.*, 2012), Attitudes towards Violence Honour Scale (Işık & Sakallı-Uğurlu, 2009), Psychological Maltreatment of Women Inventory (Boyacıoğlu *et al.*, 2020), Psychological Maltreatment of Women Inventory (Cem-Ersoy *et al.*, 2017). However, these measurements include men's attitudes and behaviours related to violence against women, which is a sub-dimension of sexism. On the other hand, the SGEMS (Sudkamper *et al.*, 2020) measures how men contribute to the division of labor within the household and gender equality in public spaces; it doesn't include attitudes and behaviors towards violence against women's movements or tendencies.

Another group of measures focuses on measuring men's perceptions of gender roles are Masculine Gender Role Stress Scale (Bayar *et al.*, 2018), the Perceived Threat to Manhood Scale (Türkoğlu, 2013), and the Gender Roles Attitude Scale (Zeyneloğlu & Terzioğlu, 2011). In these studies, men's attitudes and perceptions towards gender roles are prioritised rather than the steps to be taken for gender equality and behavioral dimensions. These existing scales may provide preliminary information about men's participation in gender equality, but they do not measure the extent to which they participate.

## 1.3. Current Study

This study aims to adapt the SGEMS (Sudkamper *et al.*, 2020) to Turkish, a scale originally developed with heterosexual men in North America and the UK. The confirmatory factor analysis stated that the factor structure that best represents SGEMS consists of participation in gender equality in public and domestic places. Public Domain Sub-scale measures men's intentions and behaviors regarding gender equality in public and societal contexts such as work, politics, and public decision-making. Domestic Domain Sub-scale focuses on men's behavioral intentions and actions within the household, specifically related to the division of labor and involvement in domestic responsibilities such as housework and childcare. Together, these sub-scales assess men's participation in gender equality across public and domestic spheres.

In this context, one of the possible results expected in this study is as follows:

1. As in the original scale, the scale is expected to consist of two sub-dimensions in adapting the scale to the Turkish culture. In these sub-dimensions, it is expected that the first nine items in the scale will be grouped to form the sub-dimension of participation in gender

equality in the public sphere, and the remaining seven items will be grouped to form the sub-dimension of participation in gender equality in the household.

2. The overall score of the SGEMS, the domestic sub-dimension, and the public sub-dimension scores show a negative and statistically significant correlation with the Hostile Sexism sub-dimension and the Benevolent Sexism sub-dimension of ASI.

In the original development study, the Ambivalent Sexism Inventory (ASI: Glick & Fiske, 1996; Sakallı-Uğurlu, 2002) was one of the main measurement tools used for criterion validity. The scale comprises two subscales: Hostile Sexism reflects negative, antagonistic attitudes towards women, emphasizing beliefs in women's inferiority and their roles as manipulative or demanding. Benevolent Sexism, while seemingly positive, encompasses patronizing attitudes towards women that reinforce traditional gender roles. It suggests that women are deserving of protection and care, reinforcing gender inequality through idealized, yet restrictive, perceptions of femininity. The SGEMS general, domestic, and public subscales showed a moderate negative correlation with Hostile Sexism, while correlations with Benevolent Sexism were negative but not statistically significant. As this study adapts the scale to Turkish culture, different results may emerge.

3. There will be a positive and statistically significant correlation between the scores obtained from each sub-dimension of the GEMS and the scores of participation in gender equality in the household of the SGEMS.

The Gender Equitable Men Scale (GEMS: Pulerwitz & Barker, 2008; Uçan & Baydur, 2016) was administered alongside the SGEMS. While both scales address gender-related issues, they differ fundamentally: the GEMS focuses on attitudes justifying violence against women and stereotypes about female roles, such as 'women should obey their husbands in all matters,' rather than measuring men's active participation in gender equality. In contrast, the SGEMS directly assesses men's involvement in fostering equality, both at home (e.g., 'I make all important decisions together with my partner') and in public spaces. Thus, while the GEMS complements the SGEMS in validation, it is insufficient to fully capture men's participation in gender equality across private and public domains.

The pre-registration of the study can be viewed on the OSF (Open Science Framework) page: <https://osf.io/9bu8g>.

## 2. METHOD

### 2.1. Participants

The sample consisted of only cisgender, heterosexual men over 18 from Turkey, selected using criterion sampling (Büyüköztürk *et al.*, 2008), as the original scale was developed for this group (Sudkamper *et al.*, 2020), and social psychological experiences differ between cisgender and trans men (American Psychological Association, 2015; Morgenroth & Ryan, 2018; Tate *et al.*, 2014). The study focused on heterosexual men, as the 'domestic' subscale is more relevant to those in romantic relationships with women (Sudkamper *et al.*, 2020). Data from participants under 18 or those who did not complete at least one item of the adapted SGEMS were excluded.

Following the recommendations of Catell (1978), Everitt (1975), and Kline (2013), the minimum number of participants per item was set at 20, resulting in a required sample size of 320 for the 16-item scale adaptation. The estimated effect size and power were based on the original scale development (Sudkamper *et al.*, 2020). Data were collected from 453 participants for the SGEMS adaptation study. Detailed information about the participants is provided in Table 1. Missing data and outlier analyses were conducted to prepare the data for analysis. Data from 11 participants who left all items unanswered were excluded. Multivariate normality was checked using Mahalanobis distance, revealing 23 outliers at the  $p = .001$  significance level, which were also excluded. The final analysis included data from 419 participants, ages 18 to 70 ( $M = 31.8$ ,  $SD = 11.2$ ).

**Table 1.** *Participants' relationship, divorce, and household sharing statutes.*

		N	%
Relationship status	legally married.	152	36.2
	not legally married, but have a partner.	125	29.8
	neither married nor in a partnership.	141	34
Have you been divorced before	yes	16	4.7
	no	403	95.3
Household sharing status	living with spouse	142	33.5
	living with unmarried partner	9	2.3
	living with unrelated female housemates	3	0.9
	living with unrelated male housemates	26	6.3
	living with unrelated male and female housemates	1	0.5
	living with family	149	35.4
	living alone	89	21.1

## 2.2. Data Collection Tools

### 2.2.1. Support for gender equality amongst men scale

The SGEMS, developed by Sudkamper *et al.* (2020), includes 16 items rated on a 7-point Likert scale (1 represents "completely disagree" and 7 "completely agree."), with higher scores indicating greater support for gender equality. Their CFA with heterosexual cis-men from North America and the UK revealed a two-factor structure: public and domestic support.

The scale was translated into Turkish for cultural adaptation. The authors independently translated the scale, and the translations were compared for technical accuracy, word choice, readability, and comprehension. The Turkish version was then translated into English by two individuals: one a PhD student in Psychology at a UK university, and the other a lecturer with a PhD from a UK university, now working at a Turkish university. Two Turkish language experts were included to resolve discrepancies and finalize the scale items. The finalized Turkish version of the scale is available in the [Appendix 1](#).

### 2.2.2. Ambivalent sexism inventory

The ASI (Glick & Fiske, 1996; Sakallı-Uğurlu, 2002) was administered alongside the SGEMS as a criterion measure. It includes 22 items across two sub-dimensions: hostile sexism (e.g., "Women exaggerate problems in the workplace") and benevolent sexism (e.g., "Women should be cherished and protected by men"). Participants rated items on a 6-point Likert scale (1 = strongly disagree, 6 = strongly agree), with higher scores indicating greater sexism.

The scale's adaptation study assessed validity through factor analysis and criterion-based validity by examining correlations with similar scales. Reliability was measured using Cronbach's alpha (.85) and test-retest correlation (.87). The ASI showed a .60 correlation with Burt's (1980) sex-role stereotyping measure, and its original factor structure was confirmed in the Turkish version.

### 2.2.3. The gender equality in Turkish men scale

The Gender Equitable Men Scale, first developed by Pulerwitz and Barker (2008) and updated by Nanda (2011), was adapted to Turkish by Uçan and Baydur (2016). The scale aims to measure men's attitudes towards gender inequality. The GEMS includes four sub-dimensions and 23 items in total. These are: domestic violence (6 items), assessing attitudes toward violence against wives/partners; sexual relationships (7 items), reflecting beliefs about male dominance in sexual matters; health and disease prevention (5 items), covering stigmatizing views on

sexual health; and household chores (5 items), addressing domestic roles and decision-making. Items are rated on a 3-point Likert scale (1 = *agree*, 2 = *somewhat agree*, 3 = *disagree*), with higher scores indicating greater support for gender equality.

Confirmatory factor analysis (CFA) was conducted to assess validity, confirming the original scale structure. The goodness-of-fit indices indicated an adequate model fit. The correlation of the scale with similar measures and the Cronbach's alpha coefficient were examined to evaluate reliability. The results revealed a moderate to strong correlation between the GEMS and other gender-related scales ( $p < 0.05$ ). The overall Cronbach's alpha coefficient for the scale was .85, while the sub-dimension coefficients ranged from .41 to .78.

### 2.3. Data Collection

Ethics committee approval was obtained from Anadolu University Social Sciences and Humanities Scientific Research and Publication Ethics Committee. The participants were reached via social media and answered the demographic questionnaire, SGEMS, ASI, and GEMS, respectively. The analyses were conducted using SPSS and the Mplus package programmes. The data and materials of the study can be accessed from the following link: [https://osf.io/rvamu/?view\\_only=eb79d4a20ecc40e5bf3596ee3c4d7b7d](https://osf.io/rvamu/?view_only=eb79d4a20ecc40e5bf3596ee3c4d7b7d).

### 2.4. Data Analysis

For the Turkish adaptation of the SGEMS, a CFA was conducted to assess validity and test the fit of the original scale's two-factor structure (Hypothesis 1). In the estimations, since it was determined that the data met the multivariate normality condition as a result of the assumption checks, the Maximum Likelihood (ML) method, which is widely preferred in parameter estimation for data that meet the normality condition, was preferred (Kline, 2019). In order to evaluate the fit of the data to the predicted structure, comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square error of squares (SRMR) values were examined (Kline, 2019). CFI and TLI values  $\geq .90$  and RMSEA and SRMR values  $\leq .08$  were considered as acceptable goodness values (Hu & Bentler, 1999; Schermelleh-Engel *et al.*, 2003; Tabachnick & Fidell, 2001).

For the validity study, the correlations between the scales measuring similar constructs and the data obtained from SGEMS were examined (Hypothesis 2 and Hypothesis 3). For the reliability study of the scale, internal consistency coefficients were calculated for the whole scale and the sub-dimensions. Since Cronbach's  $\alpha$  coefficient shows bias in congeneric measurements, it is recommended to calculate McDonald's Omega ( $\omega$ ) coefficient, which produces more consistent results (Zinbarg *et al.*, 2005). In this context, both Cronbach's  $\alpha$  and McDonald's Omega ( $\omega$ ) coefficients were reported for this scale.

Finally, to prove that the scale measures the same construct in different groups, the participants were divided into two groups using the 'age' information obtained from the participants in the demographic section. The age groups are those younger than 29 years old and those who are 29 years old and above. Emerging adulthood is a developmental stage proposed by Arnett (2000), typically spanning from ages 18 to 29. This phase is characterized by significant exploration in identity, relationships, career, and lifestyle choices, as individuals transition from adolescence to full adulthood (Arnett, 2000; Nelson & Barry, 2005). After the age of 29, people are considered to have entered adulthood. The multiple-group CFA measurement invariance model was tested to examine whether the scale shows measurement invariance for these two groups. For these two groups, it was gradually checked whether configural, metric, and scalar invariance was achieved. Brown (2015) states that in invariance checks with multiple group CFA, the CFA model should first be tested separately in each group. The CFA model was first tested separately for emerging adults (18-29) and adult (30 and over) groups, and the goodness of fit values were examined. Then, the measurement invariance steps were tested step by step. In the controls of the invariance steps, the observed change in chi-square, CFI, and RMSEA

values was analyzed. For measurement invariance, the observed change in chi-square values ( $\Delta\chi^2$ ) should be insignificant ( $p > .05$ ), the observed change in CFI values ( $\Delta\text{CFI}$ ) should be less than .01, and the observed change in RMSEA values ( $\Delta\text{RMSEA}$ ) should be less than .015 (Cheung & Rensvold, 2002).

The data set must meet certain assumptions for the validity and reliability analyses. In this context, multivariate and univariate normality, multicollinearity, multicollinearity and linearity checks of the data set were performed first to prove the relevant analyses could be used. For assumption controls, descriptive statistics, skewness (-0.789), and kurtosis (0.934) coefficients of the data set calculated to examine normality were examined, and it was seen that these values were within the range of ( $\pm 1$ ). In addition, kurtosis and skewness also showed that histogram graphs were close to a normal distribution, and Q-Q Plot graphs followed a linear distribution, and it was accepted that the data showed a normal distribution with no significant deviation (Tabachnick & Fidell, 2001). Bartlett's test of sphericity was performed for the multivariate normality assumption of the data, and it was determined that the multivariate normality condition was also met ( $\chi^2_{(120)} = 3236.84, p < .001$ ). Multicollinearity checks of the data set and inter-item correlation values were calculated, and these values are presented in Table 2.

**Table 2.** Inter-item correlation coefficients.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	0.44	0.58	0.44	0.41	0.44	0.47	0.46	0.39	0.31	0.23	0.30	0.33	0.29	0.14	0.27
2		1	0.63	0.45	0.38	0.37	0.44	0.32	0.37	0.10	0.15	0.19	0.28	0.25	0.26	0.17
3			1	0.61	0.40	0.47	0.50	0.41	0.45	0.23	0.18	0.23	0.30	0.32	0.19	0.23
4				1	0.33	0.39	0.41	0.40	0.42	0.26	0.19	0.24	0.30	0.29	0.23	0.18
5					1	0.73	0.58	0.45	0.44	0.22	0.18	0.30	0.36	0.30	0.16	0.21
6						1	0.60	0.51	0.46	0.25	0.23	0.31	0.35	0.32	0.19	0.24
7							1	0.62	0.59	0.33	0.22	0.31	0.33	0.33	0.17	0.23
8								1	0.64	0.40	0.24	0.35	0.35	0.34	0.18	0.26
9									1	0.33	0.26	0.28	0.36	0.34	0.24	0.24
10										1	0.32	0.25	0.32	0.34	0.20	0.19
11											1	0.49	0.41	0.36	0.26	0.22
12												1	0.59	0.48	0.26	0.31
13													1	0.76	0.37	0.21
14														1	0.46	0.22
15															1	0.18
16																1

Table 2 shows that the correlation values vary between .14 - .76. There is no correlation value exceeding .90 in the data set; generally, the variables show moderate correlation with each other. Accordingly, there is no multicollinearity problem in the data set.

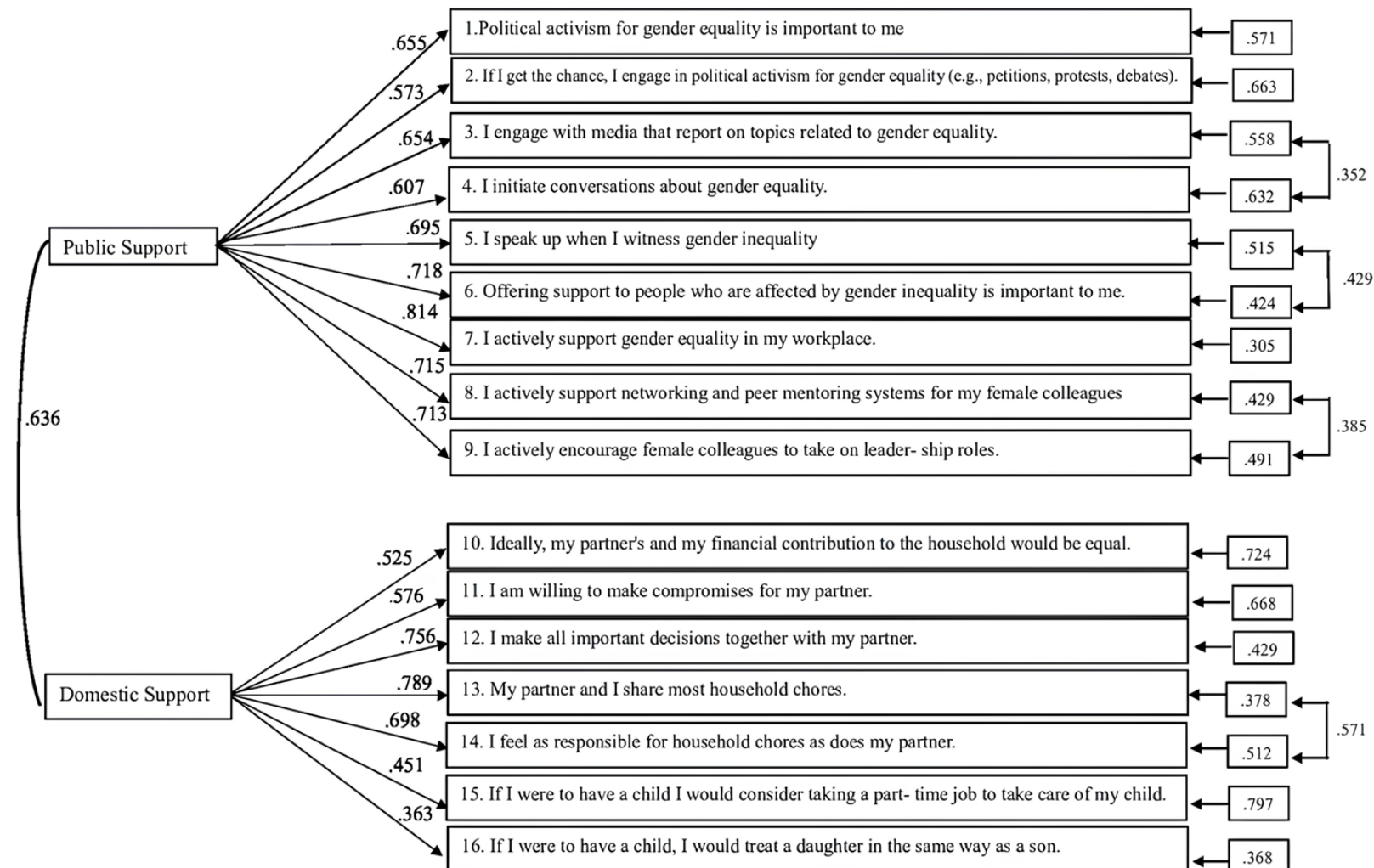
### 3. RESULTS

#### 3.1. Results Related to Validity

##### 3.1.1. Confirmatory factor analysis

As a result of the CFA conducted to confirm the two-factor structure that emerged in the original scale, the first goodness of fit values obtained for the model were estimated as  $\chi^2/df = 599/103 = 5.81, p < .001$ , CFI = .84, TLI = .82, RMSEA = .11 (90% CI: .09, .12), SRMR = .07. When these values were analyzed, it was seen that except SRMR, the other indices were not within the acceptable limits of goodness of model fit ( $\chi^2/df < 3$ , TLI and CFI  $> .09$ , RMSEA and SRMR  $< .08$ ) (Kenny et al., 2015).



**Figure 1.** CFA diagram for SGEMS.

In this context, modification suggestions were analyzed. While selecting the ones to be applied among the modification proposals, the proposals with the highest MI values, the scale's original structure, and the experts' logical evaluations about the item contents were all examined together. In this context, it was decided to implement the 3 proposals with the highest MI values. These suggestions are the suggestions for defining a connection between items 3-item 4, item 5-item 6, items 13-item 14, and item 9-item 8 residuals. When the CFA application performed in the development of the scale was examined, it was seen that the same connections were established in the original development study of the scale. On the other hand, when the item contents were analyzed, it was seen that these item pairs were intended to measure quite similar variables. When all these analyses were evaluated holistically, it was decided to apply these modifications. The goodness of fit indices obtained after the modifications are as follows:  $\chi^2/df = 328/99 = 3.31$ ,  $p < .001$ , CFI = .928, TLI = .913, RMSEA = .07 (90% CI: .06, .08), SRMR = .074. The coefficients of the tested model are presented in Table 3, and the CFA diagram is presented in Figure 1.

**Table 3.** SGEMS factor loadings.

Items	<i>M</i>	<i>SD</i>	Item-Total Correlation	Public Support	Domestic Support
1	5.43	1.57	.66**	.65	
2	3.37	1.98	.62**	.58	
3	4.10	1.84	.70**	.66	
4	3.92	1.84	.64**	.61	
5	5.40	1.49	.65**	.69	
6	5.45	1.47	.69**	.72	
7	5.42	1.59	.72**	.83	
8	5.43	1.42	.69**	.71	
9	5.41	1.59	.69**	.71	
10	5.33	1.73	.52**	.31	.52
11	5.17	1.55	.49**		.58
12	5.76	1.28	.57**		.76
13	5.56	1.47	.66**		.79
14	5.65	1.51	.64**		.70
15	4.47	1.97	.47**		.45
16	6.25	1.39	.43**		.36
Eigenvalue				6.24	1.79
Explained variance (%)				38.99	11.22
Cronbach's $\alpha$				.89	.78
McDonald's Omega				.90	.81

### 3.1.2. Criterion-based validity findings

The study aimed to assess the criterion-related validity of the SGEMS by examining the correlations between SGEMS scores and scores from the ASI and GEMS, both for sub-dimensions and total scores. Based on the constructs measured by the scales, negative significant correlations were expected between SGEMS and ASI scores, and positive significant correlations with GEMS. The correlation coefficients for the total scores and sub-dimensions are provided in Table 4.

As expected, the total score obtained from the SGEM scale has a significant negative correlation with the total score obtained from the ASI. Similarly, the hostile sexism sub-dimension of the ESLS has a significant negative correlation with the total score obtained from the ECEK scale and the scores obtained from the public and domestic sub-dimensions. These correlations indicate low-strength relationships ranging between  $-.17$  and  $-.31$ . On the other hand, benevolent sexism showed a negative correlation with the SGEMS total score and the public sub-dimension, as expected, while it showed an insignificant but positive correlation with domestic participation in gender equality. However, since the correlation coefficients range between  $-.07$  and  $.03$ , there is a weak relationship between protectionist sexism and the total and sub-dimensional scores of the SGEM scale.

Significant and positive relationships exist between the total scores obtained from the SGEM scale and GEMS and the scores obtained from the sub-dimensions of both scales. The correlations between the SGEMS total score and the total score obtained from the GEMS and the scores obtained from the sub-dimensions of the GEMS scale vary between  $.24$  and  $.43$ , corresponding to a moderate strength of relationship. The correlations between the SGEMS public sub-dimension and the total and sub-dimensions of the GEMS also show that there are moderate relationships (ranging between  $.26$  and  $.43$ ). When the total scores and sub-dimensions of SGEMS in domestic sub-dimension and GEMS were analyzed, weak positive significant correlations ranging between  $.14$  and  $.30$  were found.

### 3.2. Results Related to Reliability

Internal consistency coefficients were calculated to obtain reliability evidence for the SGEM Scale. In the study, two different internal consistency coefficients (Cronbach's  $\alpha$  and McDonald's Omega ( $\omega$ )) were reported. Cronbach's  $\alpha$  coefficient was reported in the report in which the scale was developed, and Cronbach's  $\alpha$  coefficient was reported for this adaptation study in order to make comparative discussions, on the other hand, since Cronbach's  $\alpha$  coefficient shows bias in congeneric measurements in the literature, it is recommended to calculate McDonald's Omega ( $\omega$ ) coefficient, which produces more consistent results (Yurdugül, 2006; Zinbarg *et al.*, 2005). In this context, McDonald's Omega ( $\omega$ ) coefficient was reported in addition to Cronbach's  $\alpha$  coefficient.

When Table 3 is analyzed, it is seen that Cronbach's  $\alpha$  coefficient is between  $.78$ -. $90$  and McDonald's Omega ( $\omega$ ) coefficient is between  $.81$ -. $90$ . These values obtained can be interpreted as quite high-reliability coefficients.

### 3.3. Findings Related to Measurement Invariance

One of the important pieces of evidence that should be presented in scale development/adaptation studies is the measurement invariance study to show that the scale works similarly for some subgroups. In this study, measurement invariance analyses were conducted by testing measurement invariance models with the help of multi-group CFA analyses. It was tested whether the emerging adulthood (18-29) and adulthood (30 and above), which appeared in 2 groups formed based on age variable, worked similarly. Brown (2015) states that in invariance checks with multiple group CFA, the CFA model should first be tested separately in each group. In this context, the CFA model was first tested separately for emerging adult and adult groups, and then the measurement invariance steps were tested. The goodness of fit is presented in Table 5.

Table 5 shows that the goodness of fit indices related to the model validation process are presented separately for the two groups. When the goodness of fit indices are examined, it is seen that the indices are within the desired limits ( $\chi^2/df < 3$ , TLI and CFI  $> .90$ , RMSEA and SRMR  $< .08$ ) for both groups. Accordingly, it can be said that the CFA model was confirmed for both groups separately.

**Table 4.** Means, standard deviations, and correlation coefficients for the factors of the SGEMS and related variables.

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11
1. SGEMS Total	5.13	0.99	1	.92**	.82*	-.17*	-.14*	-.18**	.41**	.24**	.34**	.33**	.42**
2. SGEMS Public	4.88	1.19		1	.53**	-.22**	-.18**	-.23**	.43**	.26**	.34**	.35**	.42**
3. SGEMS Domestic	5.45	1.03			1	-.04	-.04	-.05	.26**	.14*	.22**	.20**	.30**
4. ASI Total	81.76	20.22				1	.97**	.97**	-.49**	-.27**	-.52**	-.36**	-.33**
5. ASI Benevolent Sexism	41.95	10.58					1	.87**	-.48**	-.26**	-.51**	-.34**	-.33**
6. ASI Hostile Sexism	39.79	10.41						1	-.47**	-.29**	-.50**	-.35**	-.31**
7. GEMS Total	65.99	6.44							1	.74**	.88**	.80**	.81**
8. GEMS Domestic Violence	17.29	1.46								1	.49**	.52**	.53**
9. GEMS Sexual Relationships	20.82	3.00									1	.60**	.58**
10. GEMS Health and Disease Prevention	13.98	1.52										1	.59**
11. GEMS Household Chores	13.88	1.84											1

Note. \* represents significance at .01 and \*\* at .001 level. SGEMS: Support for Gender Equality Amongst Men, ASI: Ambivalent Sexism Inventory, GEMS: Gender Equitable Men Scale.

**Table 5.** Goodness of fit indices for measurement invariance.

	$\chi^2$	<i>df</i>	RMSEA	SRMR	CFI	TLI	$\Delta\chi^2$	$\Delta df$	<i>p</i>	$\Delta$ RMSEA	$\Delta$ CFI	$\Delta$ TLI
Single Group Solution												
Emerging Adulthood ( <i>n</i> = 218)	194.95	98	.067	.058	.941	.928						
Adulthood ( <i>n</i> = 184)	211.49	98	.091	.066	.909	.889						
Measurement Invariance												
Configural	406.44	196	.073	.062	.927	.911	-	-	-	-	-	-
Metric	423.13	210	.071	.070	.927	.916	16.69	14	0.273	.002	.000	.005
Scalar	455.15	224	.072	.073	.920	.915	32.02	14	0.004	.001	.007	.001

For measurement invariance, the observed change in chi-square, CFI and RMSEA values for each step was analyzed. For measurement invariance, the observed change in chi-square values ( $\Delta\chi^2$ ) should be insignificant ( $p > .05$ ), the observed change in CFI values ( $\Delta\text{CFI}$ ) should be less than .01 and the observed change in RMSEA values ( $\Delta\text{RMSEA}$ ) should be less than .015 (Cheung & Rensvold, 2002). In this sense, when the  $\Delta\chi^2$  and  $p$  values in Table 5 are analyzed, it is seen that the scale provides Configural and Metric change ( $p > .05$ ) but the scalar invariance step is not provided ( $p < .05$ ). Another evaluation criterion when examining measurement invariance is to look at the difference between CFI and RMSEA. On the other hand, it is stated that  $\chi^2$  value is not a useful criterion for model fit because it is sensitive to sample size (Cheung & Rensvold, 2002). In this context, it is seen that  $\Delta\text{CFI} < .01$  and  $\Delta\text{RMSEA} < .015$  in metric and scalar invariance steps. When the CFI and RMSEA values are taken into consideration, it can be said that the scale provides configural, metric and scalar invariance steps.

#### 4. DISCUSSION and CONCLUSION

Gender equality studies typically focus on women's empowerment and the steps they can take to achieve equality (Ryan & Kirby, 2018). However, it is crucial to consider men's participation in these efforts (Cihangir *et al.*, 2014; Drury & Kaiser, 2014). Existing measures of men's engagement with gender equality are insufficient. Nevertheless, this issue is partially addressed by the Support for Gender Equality among Men Scale (SGEMS: Sudkamper *et al.*, 2020). This study adapts the SGEMS to Turkish, providing evidence of its validity and reliability in Turkey. The factor analysis confirmed the construct validity of the SGEMS, aligning with the original study's two-factor structure: participation in gender equality in public and domestic domains (Sudkamper *et al.*, 2020).

In this study, both SGEMS sub-dimensions showed reliability above .70 (Cronbach's  $\alpha$ ), meeting the threshold for reliability (DeVellis, 1991; Kline, 1986). However, the domestic sub-dimension ( $\alpha = .89$ ) had lower reliability than the public sub-dimension ( $\alpha = .89$ ), consistent with the original scale development. To ensure accuracy, McDonald's Omega ( $\omega$ ) was also used, as it is considered a more robust reliability measure than Cronbach's  $\alpha$ , particularly after CFA (Hayes & Coutts, 2020; Sijtsma, 2009). Using  $\omega$ , the reliability gap between the public and domestic subscales was minimized.

The lower reliability for the domestic subscale may be linked to participants' demographics. Specifically, 33.7% were single, 95.3% had never married, 6.3% lived with only male housemates, and 21.1% lived alone, potentially leading them to answer household-related questions hypothetically. Additionally, 35.4% lived with family members, often in traditionally patriarchal households (Bozok, 2018), limiting their opportunities to practice or advocate for gender equality at home.

The domestic sub-dimension of the SGEMS is a crucial tool for assessing men's involvement in gender equality within the Turkish context. While research on gender roles in Turkey has primarily focused on women's perspectives (Yüksel-Kaptanoğlu & Çavlin, 2015), with studies showing that 71% of married women believe men should share housework equally and 65% support their active role in childcare, the SGEMS provides a valuable means to directly evaluate men's attitudes and behaviors regarding domestic responsibilities. This shift in focus is essential for promoting men's active participation in achieving gender equality in household spheres.

Criterion-referenced validity was evaluated by examining correlations between SGEMS, ASI, and GEMS scores. As expected, SGEMS total and subscale scores showed significant negative correlations with ASI, particularly its hostile sexism sub-dimension, aligning with the original study and literature. However, no significant correlation was found between benevolent sexism and SGEMS scores. While benevolent sexism was negatively associated with the total SGEMS score and public sub-dimension, it showed a weak, insignificant positive link with the household sub-dimension. This aligns with Sudkamper *et al.* (2020) and Glick and Fiske's (1996, 2001) argument that men may not recognize benevolent sexism as discriminatory, often



viewing protective behaviors as supportive. For example, actions like compromising to reconcile with a partner may reinforce traditional roles rather than promote equality. The weak negative correlation with the public subscale suggests men with high benevolent sexism may perceive public gender equality efforts as personal sacrifices rather than shared responsibilities.

Another possible explanation for the lack of a significant relationship between benevolent sexism and SGEMS total and subscale scores is the tendency of individuals from advantaged groups to feel a sense of responsibility toward disadvantaged groups, a phenomenon known as *noblesse oblige* (Fiddick & Cummins, 2007; Vanbeselaere et al., 2006). In a similar vein, Glick et al. (2004) suggest that benevolent sexism allows men to present themselves as supportive of gender equality while maintaining their societal advantages, making sexism more implicit in contemporary contexts. Consequently, in this study, men with higher benevolent sexism scores may still appear to support gender equality despite underlying biases.

A positive and significant correlation between GEMS and SGEMS was expected, and the findings support this expectation. All sub-dimensions of both scales exhibit significant positive relationships. However, unlike GEMS, which primarily assesses negative attitudes toward women, SGEMS specifically measures men's participation in gender equality. For instance, GEMS includes items reflecting acceptance of violence against women (e.g., *"There are times when a woman deserves to be beaten"*). In contrast, SGEMS focuses on positive behaviors (e.g., *"I make all important decisions together with my partner"*). These differences highlight the necessity of adapting SGEMS to Turkish, as it captures men's active participation in gender equality rather than merely assessing negative attitudes toward women. While GEMS provides insight into sexist beliefs, SGEMS offers a complementary perspective by focusing on behavioral intentions. Therefore, the adaptation of SGEMS contributes to a more comprehensive assessment of men's engagement with gender equality in Turkey.

A potential limitation is the high proportion of partnered participants (over 60%), as men in partnerships often show greater support for gender equality. Such relationships can increase awareness of gender roles, promote empathy, and reinforce egalitarian values through shared experiences and mutual influence (Olliffe, Kelly, Gonzales et al., 2022). While initial findings support the SGEMS as a reliable tool for assessing men's support for gender equality, further research is needed to confirm its factor structure across diverse populations and cultural contexts in Turkey. In regions with limited gender equality, some items—particularly those on inclusive workplace practices—may require adaptation to reflect local socio-cultural conditions.

The findings indicate that the Turkish adaptation of the SGEMS is a valid and reliable measure of men's behavioral intentions to engage in gender equality in both domestic and public domains. Consistent with the original scale by Sudkamper et al. (2020), a two-factor structure emerged. Unlike existing gender equality scales in Turkey, SGEMS assesses men's participation, contributing to contemporary gender research.

### Acknowledgments

We thank Dr. Antonia Sudkämper and her team for allowing us to adapt their scale and for their kind support throughout the process.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Anadolu University Social Sciences and Humanities Scientific Research and Publication Ethics Committee, ID: 269759.

### Contribution of Authors

**Esma Esen Çiftçi:** Investigation, Resources, Formal analysis, Methodology, Writing-original draft, Revisions. **Esra Daşçı:** Investigation, Resources, Methodology, Writing-original draft,

Revisions. **Cansu Ayan:** Investigation, Resources, Methodology, Formal analysis, Revisions. **Zeynep Uludağ:** Investigation, Resources, Methodology, Writing-original draft, Revisions.

### Orcid

Esma Esen Çiftçi  <https://orcid.org/0000-0002-3545-0998>

Esra Daşçı  <https://orcid.org/0000-0002-0124-9380>

Cansu Ayan  <https://orcid.org/0000-0002-0773-5486>

Zeynep Uludağ  <https://orcid.org/0000-0002-0447-2158>

### REFERENCES

- Achen, A.C., & Stafford, F.P. (2005). *Data quality of housework hours in the panel study of income dynamics: Who really does the dishes*. University of Michigan.
- Ajzen, I. (1991). The theory of planned behavior. *Organisational Behavior and Human Decision Processes*, 50, 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Ajzen, I., & Sheikh, S. (2013). Action versus inaction: Anticipated affect in the theory of planned behavior. *Journal of Applied Social Psychology*, 43, 155-162. <https://doi.org/10.1111/j.1559-1816.2012.00989.x>
- Akın, A., Gülşen, M., Aşut, S., & Akca, M. (2012). Yakın İlişkilerde Şiddet Sorumluluğu Ölçeği Türkçe formunun geçerlik ve güvenirliği [Validity and reliability of the Turkish version of the Intimate Partner Violence Responsibility Scale]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 12(2), 175-184.
- American Psychological Association. (2015). Guidelines for psychological practice with transgender and gender nonconforming people. *American Psychologist*, 70(9), 832–864. <https://doi.org/10.1037/a0039906>
- Anderson, V.N. (2009). What's in a label? Judgments of feminist men and feminist women. *Psychology of Women Quarterly*, 33, 206-215. <https://doi.org/10.1111/j.1471-6402.2009.01490.x>
- Arnett, J.J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, 55(5), 469-480. <https://doi.org/10.1037/0003-066X.55.5.469>
- Bayar, Ö., Haskan Avcı, Ö., & Koç, M. (2018). Erkek Toplumsal Cinsiyet Rolü Stresi Ölçeği'nin (ETCRSÖ) geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of the Male Gender Role Stress Scale (MGRSS): A validity and reliability study]. *Abant İzzet Baysal Üniversitesi Fakültesi Dergisi*, 18(1), 57-76.
- Bear, J.B., Treviño, L.J., & Aguinis, H. (2025). A star is born or not: Understanding the star emergence gender gap. *Journal of Organizational Behavior*, 46, 351-367. <https://doi.org/10.1002/job.2858>
- Belingheri, P., Chiarello, F., Fronzetti Colladon, A., & Rovelli, P. (2021) Twenty years of gender equality research: A scoping review based on a new semantic indicator. *PLoS ONE* 16(9). <https://doi.org/10.1371/journal.pone.0256474>
- Bianchi, S.M., Milkie, M.A., Sayer, L.C., & Robinson, J.P. (2000). Is anyone doing the housework? Trends in the gender division of household labor. *Social Forces*, 79, 191-228.
- Boyacıoğlu, İ., Uysal, M.S., & Erduran, C. (2020). Psikolojik şiddetin ölçümü: Psikolojik İstismar Profiline ve Kadına Kötü Muamele Envanterinin Türkçe'ye uyarlanması [Measuring psychological violence: Adaptation of the Psychological Abuse Profile and the Inventory of Violence Against Women into Turkish]. *Psikoloji Çalışmaları*, 40(1), 19-55. <https://doi.org/10.26650/SP2019-0027>
- Branscombe, N.R., & Deaux, K. (1991). Feminist attitude accessibility and behavioral intentions. *Psychology of Women Quarterly*, 15(3), 411-418. <https://doi.org/10.1111/j.1471-6402.1991.tb00417.x>

- Breinlinger, S., & Kelly, C. (1994). Women's responses to status inequality: A test of social identity theory. *Psychology of Women Quarterly*, 18(1), 1-16. <https://doi.org/10.1111/j.1471-6402.1994.tb00293.x>
- Brown, M.J., & Gladstone, N. (2012). Development of a short version of the gender role beliefs scale. *International Journal of Psychology and Behavioral Sciences*, 2, 154-158. <https://doi.org/10.5923/j.ijpbs.20120205.05>
- Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2<sup>nd</sup> ed.). Guilford Publications.
- Burt, M.R. (1980). *Sex Role Stereotyping Scale* [Database record]. APA PsycTests. <https://doi.org/10.1037/t06663-000>
- Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., & Karadeniz, Ş. (2008). *Bilimsel araştırma yöntemleri [Scientific research methods]*. Pegem Akademik Yayıncılık.
- Bozok, M. (2018). *Ebeveynlik, erkeklik ve çalışma hayatı arasında Türkiye’de babalık [Fatherhood in Turkey: Between parenthood, masculinity, and work life]*. AÇEV. <https://dspace.ceid.org.tr/handle/1/801>
- Catell, R.B. (1978). *The scientific use of factor analysis*. Plenum.
- Cem-Ersoy, N., Hünler, O.S., & Namer, Y. (2017). Kadına Psikolojik Eziyet Envanteri kısa formu Türkçe uyarlaması [Turkish adaptation of the short form of the Psychological Maltreatment of Women Inventory]. *Klinik Psikiyatri*, 20, 276-286.
- Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, 9(2), 233-255.
- Cho, E., Allen, T.D., & Meier, L.L. (2025). Is ‘me-time’selfish?: Daily vitality crossover in dual-earner couples. *Applied Psychology: Health and Well-Being*, 17(1). <https://doi.org/10.1111/aphw.70004>
- Cihangir, S., Barreto, M., & Ellemers, N. (2014). Men as allies against sexism: The positive effects of a suggestion of sexism by male (vs. female) sources. *Sage Open*, 4, 1-12. <https://doi.org/10.1177/2158244014539168>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> ed.). Lawrence Erlbaum Associates, Publishers.
- Craig, L., Perales, F., Vidal, S., & Baxter, J. (2016). Domestic outsourcing, housework time, and subjective time pressure: New insights from longitudinal data. *Journal of Marriage and Family*, 78, 1224-1236. <https://doi.org/10.1111/jomf.12321>
- Deutsch, F. (1999). *Halving it all: How equally shared parenting works*. Harvard University Press.
- DeVellis, R.F. (1991). *Scale development: Theory and applications*. Sage Publications, Inc.
- Dotti Sani, G.M. (2014). Men's employment hours and time on domestic chores in European countries. *Journal of Family Issues*, 35, 1023-1047. <https://doi.org/10.1177/0192513x14522245>
- Drury, B.J., & Kaiser, C.R. (2014). Allies against sexism: The role of men in confronting sexism. *The Journal of Social Issues*, 70, 637-652. <https://doi.org/10.1111/josi.12083>
- Eagly, A.H., & Karau, S.J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573-598. <https://doi.org/10.1037/0033-295X.109.3.573>
- Eliezer, D., & Major, B. (2011). It's not your fault: The social costs of claiming discrimination on behalf of someone else. *Group Processes & Intergroup Relations*, 15, 487-502. <https://doi.org/10.1177/1368430211432894>
- Estevan-Reina, L., de Lemus, S., & Megías, J.L. (2017, June). *Can men be allies of women in fight against sexism? Feminist identity and benevolent sexism as predictors of sexism confrontation for paternalistic vs. egalitarian reasons* [Poster presentation]. EASP: Gender Roles in the Future? Theoretical Foundations and Future Research Directions, Berlin, Germany.

- Everitt, B.S. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, 126(3), 237-240. <https://doi.org/10.1192/bjp.126.3.237>
- Fiddick, L., & Cummins, D. (2007). Are perceptions of fairness relationship-specific? The case of noblesse oblige. *Quarterly Journal of Experimental Psychology*, 60(1), 16-31. <https://doi.org/10.1080/17470210600577266>
- Frei, J.R., & Shaver, P.R. (2002). Respect in close relationships: Prototype definition, self-report assessment, and initial correlates. *Personal relationships*, 9(2), 121-139.
- Gärtner, M. (2007). *FOCUS: Fostering caring masculinities*. Dissens e.V. and genderWerk.
- Glick, P., & Fiske, S. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3), 491-512. <https://doi.org/10.1037/0022-3514.70.3.491>
- Glick, P., & Fiske, S.T. (2001). An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American Psychologist*, 56(2), 109-118. <https://doi.org/10.1037/0003-066X.56.2.109>
- Glick, P., Lameiras, M., Fiske, S.T., Eckes, T., Masser, B., Volpato, C., Manganelli, A.M., Pek, J.C.X., Huang, L.-l., Sakalli-Uğurlu, N., Castro, Y.R., D'Avila Pereira, M.L., Willemsen, T.M., Brunner, A., Six-Materna, I., & Wells, R. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, 86(5), 713-728. <https://doi.org/10.1037/0022-3514.86.5.713>
- Gurin, P., & Townsend, A. (1986). Properties of gender identity and their implications for gender consciousness. *British Journal of Social Psychology*, 25(2), 139-148. <https://doi.org/10.1111/j.2044-8309.1986.tb00712.x>
- Heilman, M.E. (2012). Gender stereotypes and workplace bias. *Research in Organisational Behavior*, 32, 113-135. <https://doi.org/10.1016/j.riob.2012.11.003>
- Hendrick, S.S., & Hendrick, C. (2006). Measuring respect in close relationships. *Journal of Social and Personal Relationships*, 23(6), 881-899. <https://doi.org/10.1177/0265407506070471>
- Haas, L. (2003). Parental leave and gender equality: Lessons from the European Union. *Review of Policy Research*, 20, 89-114. <https://doi.org/10.1111/1541-1338.d01-6>
- Hayes, A.F., & Coutts, J.J. (2020). Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But.... *Communication Methods and Measures*, 14(1), 1-24. <https://doi.org/10.1080/19312458.2020.1718629>
- Hopkins-Doyle, A., Sutton, R.M., Douglas, K.M., & Calogero, R.M. (2019). Flattering to deceive: Why people misunderstand benevolent sexism. *Journal of Personality and Social Psychology*, 116(2), 167-192. <https://doi.org/10.1037/pspa0000135>
- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Işık, R., & Sakallı-Uğurlu, N. (2009). Namusa ve Namus Adına Kadına Uygulanan Şiddetle İlişkin Tutumlar Ölçeklerinin öğrenci örneklemiyle geliştirilmesi [Development of the Attitudes Toward Honor and Violence Against Women in the name of honor scales with a student sample]. *Türk Psikoloji Yazıları*, 12(24), 16-24.
- Iyer, A., & Ryan, M.K. (2009). Why do men and women challenge gender discrimination in the workplace? The role of group status and in-group identification in predicting pathways to collective action. *Journal of Social Issues*, 65, 791-814. <https://doi.org/10.1111/j.1540-4560.2009.01625.x>
- Kato-Wallace, J., Barker, G., Eads, M., & Levto, R. (2014). Global pathways to men's caregiving: Mixed methods findings from the international men and gender equality survey and the men who care study. *Global Public Health*, 9, 706-722. <https://doi.org/10.1080/17441692.2014.921829>
- Kaufman, M., & Kimmel, M. (2011). *The guy's guide to feminism*. Seal Press.



- Kelly, C., & Breinlinger, S. (1995). Identity and injustice: Exploring women's participation in collective action. *Journal of Community & Applied Social Psychology*, 5(1), 41-57. <https://doi.org/10.1002/casp.2450050104>
- Kenny, D.A., Kaniskan, B., & McCoach, D.B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486-507. <https://doi.org/10.1177/0049124114543236>
- Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. Methuen.
- Kline, R.B. (2013). Exploratory and confirmatory factor analysis In Y. Petscher & C. Schatsschneider (Eds.), *Applied quantitative analysis in the social sciences* (pp. 171- 207). Routledge.
- Kline, R.B. (2019). *Yapısal eşitlik modellemesinin ilkeleri ve uygulaması [Principles and Practice of Structural Equation Modeling]* (S. Şen, Ed. & Trans.). Nobel Akademik Yayıncılık. (Original work published 2016)
- Kosakowska-Berezecka, N., Besta, T., Adamska, K., Jaskiewicz, M., Jurek, P., & Vandello, J. A. (2016). If my masculinity is threatened, I won't support gender equality? The role of agentic self-stereotyping in restoration of manhood and perception of gender relations. *Psychology of Men & Masculinity*, 17, 274-284. <https://doi.org/10.1037/men0000016>
- Larsen, K.S., & Long, E. (1988). Attitudes toward sex-roles: Traditional or egalitarian? *Sex Roles: A Journal of Research*, 19(1-2), 1-12. <https://doi.org/10.1007/BF00292459>
- Lemaster, P., Strough, J., Stoiko, R., & DiDonato, L. (2015). To have and to do: Masculine facets of gender predict men's and women's attitudes about gender equality among college students. *Psychology of Men & Masculinity*, 16, 195. <https://doi.org/10.1037/a0036429>
- Liff, S., & Cameron, I. (1997). Changing equality cultures to move beyond “women's problems”. *Gender, Work & Organisation*, 4, 35-46. <https://doi.org/10.1111/1468-0432.00022>
- Manstead, A.S.R., & Parker, D. (1995). Evaluating and Extending the Theory of Planned Behaviour. *European Review of Social Psychology*, 6(1), 69-95. <https://doi.org/10.1080/14792779443000012>
- Mills, M.J., Culbertson, S.S., Huffman, A.H., & Connell, A.R. (2012). Assessing gender biases: Development and initial validation of the Gender Role Stereotypes Scale. *Gender in Management*, 27, 520-540. <https://doi.org/10.1108/17542411211279715>
- Morgan, B.L. (1996). Putting the feminism into feminism scales: Introduction of a Liberal Feminist Attitude and Ideology Scale. *Sex Roles: A Journal of Research*, 34(5-6), 359-390. <https://doi.org/10.1007/BF01547807>
- Morgenroth, T., & Ryan, M. (2018). Gender trouble in social psychology: How can Butler's work inform experimental social psychologists' conceptualization of gender? *Frontiers In Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.01320>
- Moser, C.E., & Branscombe, N.R. (2022). Male allies at work: Gender-Equality supportive men reduce negative underrepresentation effects among women. *Social Psychological and Personality Science*, 13(2), 372-381. <https://doi.org/10.1177/19485506211033748>
- Nanda, G. (2011). *Compendium of Gender Scales*. FHI 360/C-Change.
- Nelson, L.J., & Barry, C.M. (2005). Distinguishing features of emerging adulthood: The role of self-classification as an adult. *Journal of Adolescent Research*, 20(2), 242-262. <https://doi.org/10.1177/0743558404273074>
- Noonan, R.K. (1995). Women against the state: Political opportunities and collective action frames in Chile's transition to democracy. *Sociological Forum*, 10, 81-111. <https://doi.org/10.1007/BF02098565>
- Oliffe, J.L., Kelly, M.T., Gonzalez Montaner, G., Seidler, Z.E., Maher, B., & Rice, S.M. (2022). Men building better relationships: A scoping review. *Health Promotion Journal of Australia*, 33(1), 126-137. <https://doi.org/10.1002/hpja.463>



- Pulerwitz, J., & Barker, G. (2008). Measuring attitudes toward gender norms among young men in Brazil: Development and psychometric evaluation of the GEM Scale. *Men and Masculinities*, 10(3), 322-338. <https://doi.org/10.1177/1097184X06298778>
- Rasinski, H.M., & Czopp, A.M. (2010). The effect of target status on witnesses' reactions to confrontations of bias. *Basic and Applied Social Psychology*, 32, 8-16. <https://doi.org/10.1080/01973530903539754>
- Ryan, M., & Kirby, T. (2018, April 12). Lean in – but how? *BPS*. <https://www.bps.org.uk/psychologist/lean-how>
- Sakallı-Uğurlu, N. (2002). Çelişik Duygulu Cinsiyetçilik Ölçeği: Geçerlik ve güvenirlik çalışması [Ambivalent Sexism Inventory: Validity and reliability study]. *Türk Psikoloji Dergisi*, 17(49), 47-58.
- Sakallı, N., & Ulu, S. (2003). Evlilikte kadına yönelik şiddete ilişkin tutumlar: Çelişik duygulu cinsiyetçilik, yaş, eğitim ve gelir düzeyinin etkileri [Attitudes toward domestic violence: The effects of ambivalent sexism, age, education, and income]. *Türk Psikoloji Yazıları*, 6(11-12), 53-65.
- Schermelleh-Engel, K., Kerwer, M., & Klein, A.G. (2014). Evaluation of model fit in nonlinear multilevel structural equation modeling. *Frontiers in Psychology*, 5, 181. <https://doi.org/10.3389/fpsyg.2014.00181>
- Sijtsma K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Spence, J.T., Helmreich, R., & Stapp, J. (1973). A short version of the Attitudes toward Women Scale (AWS). *Bulletin of the Psychonomic Society*, 2(4), 219-220. <https://doi.org/10.3758/BF03329252>
- Stewart, A.L. (2016). Men's collective action willingness: Testing different theoretical models of protesting gender inequality for women and men. *Psychology of Men and Masculinity*, 18, 1-10. <https://doi.org/10.1037/men0000068>
- Streiner, D.L. (2003). Starting at the beginning: An introduction to Coefficient Alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99-103. [https://doi.org/10.1207/S15327752JPA8001\\_18](https://doi.org/10.1207/S15327752JPA8001_18)
- Sudkämper, A., Ryan, M., Kirby, T., & Morgenroth, T. (2020). A comprehensive measure of attitudes and behaviour: Development of the Support for Gender Equality among Men Scale. *European Journal of Social Psychology*, 50(2), 256-277. <https://doi.org/10.1002/ejsp.2629>
- Tabachnick, B.G., & Fidell, L.S. (2001). *Using Multivariate Statistics* (4<sup>th</sup> ed.). Allyn and Bacon.
- Tate, C.C., Youssef, C.P., & Bettergarcia, J.N. (2014). Integrating the study of transgender spectrum and cisgender experiences of self-categorization from a personality perspective. *Review of General Psychology*, 18(4), 302-312. <https://doi.org/10.1037/gpr0000019>
- Türkoğlu, B. (2013). *Violence as a way of reconstructing manhood: The role of threatened manhood and masculine ideology on violence against women* [Unpublished master's thesis]. Middle East Technical University.
- Uçan, G., & Baydur, H. (2016). Türk Erkeklerinde Toplumsal Cinsiyet Eşitliği Ölçeğinin geçerlilik ve güvenirlik çalışması [Validity and reliability study of the Gender Equality Scale among Turkish men]. *The Journal Of Academic Social Science Studies*, 6(47), 289-289. <https://doi.org/10.9761/jasss3495>
- Vanbeselaere, N., Boen, F., Van Avermaet, E., & Buelens, H. (2006). The janus face of power in intergroup contexts: A further exploration of the noblesse oblige effect. *The Journal of Social Psychology*, 146(6), 685–699. <https://doi.org/10.3200/SOCP.146.6.685-699>
- Van Laar, C., Van Rossum, A., Kosakowska-Berezecka, N., Bongiorno, R., & Block, K. (2024). MANDatory-why men need (and are needed for) gender equality progress. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1263313>

- White, A.M. (2006). Racial and gender attitudes as predictors of feminist activism among self-identified African American feminists. *Journal of Black Psychology*, 32(4), 455-478. <https://doi.org/10.1177/0095798406292469>
- Woodzicka, J.A., & LaFrance, M. (2001). Real versus imagined gender harassment. *Journal of Social Issues*, 57(1), 15-30. <https://doi.org/10.1111/0022-4537.00199>
- Yurdugül, H. (2006). The comparison of reliability coefficients in parallel, tau-equivalent, and congeneric measurements. *Ankara University Journal of Faculty of Educational Sciences*, 39(1), 15-37. [https://doi.org/10.1501/Egifak\\_00000000127](https://doi.org/10.1501/Egifak_00000000127)
- Yüksel-Kaptanoğlu, İ., & Çavlin, A. (2015). Kadına yönelik şiddet yaygınlığı [Prevalence of violence against women]. In *Türkiye’de kadına yönelik aile içi şiddet araştırması [Domestic violence against women in Türkiye]* (pp. 81-122). Hacettepe Üniversitesi Nüfus Etütleri Enstitüsü.
- Zeyneloğlu, S., & Terzioğlu, F. (2011). Toplumsal Cinsiyet Rollerini Tutum Ölçeğinin geliştirilmesi ve psikometrik özellikleri [Development and psychometric properties of the Gender Roles Attitude Scale]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 40, 409-420.
- Zinbarg, R.E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach’s  $\alpha$ , Revelle’s  $\beta$ , and McDonald’s  $\omega$  H: Their relations with each other and two alternative conceptualisations of reliability. *Psychometrika*, 70(1), 123-133. <https://doi.org/10.1007/s11336-003-0974-7>
- Zucker, A.N. (2004). Disavowing social identities: What it means when women say, “I’m not a feminist, but ...”. *Psychology of Women Quarterly*, 28(4), 423-435. <https://doi.org/10.1111/j.1471-6402.2004.00159.x>

## APPENDIX

### **Appendix 1.** *Subdimensions and items of the Support for Gender Equality Among Men Scale (SGEMS)-Turkish version.*

---

#### **Public Support for Gender Equality (Kamusal Alanda Destek)**

---

1. Cinsiyet eşitliğine yönelik politik faaliyetler benim için önemlidir.
2. Fırsatım oldukça cinsiyet eşitliğine yönelik politik faaliyetlere katılım (örneğin; imza kampanyası, protestolar, münazaralar)
3. Medyada yer alan, cinsiyet eşitliği ile ilgili yayınlarla yakından ilgilenirim.
4. Sohbet sırasında cinsiyet eşitliği ile ilgili konular açarım.
5. Cinsiyet eşitsizliğine şahit olduğumda sesimi çıkarırım.
6. Cinsiyet eşitsizliğinden etkilenen kişilere destek olmayı önemserim.
7. İşyerimde cinsiyet eşitliğini aktif olarak desteklerim.
8. Kadın iş arkadaşlarımla ağ kurma ve akran danışmanlığı sistemlerine sahip olmasını etkin bir şekilde desteklerim.
9. Kadın iş arkadaşlarımla liderlik rollerini üstlenmeleri için etkin bir şekilde cesaretlendiririm.

---

#### **Domestic Support for Gender Equality (Hane İçi destek)**

---

10. Tercihen ben ve partnerim ev ekonomisine eşit katkıda bulunmalıyız.
  11. Partnerimle uzlaşmak için ödün vermeye istekliyimdir.
  12. Bütün önemli kararları partnerimle birlikte alırım.
  13. Partnerim ve ben ev işlerinin pek çoğunu paylaşıyoruz.
  14. Ev işlerinde kendimi partnerim kadar sorumlu hissedirim.
  15. Çocuğum olsaydı, ona bakmak için yarı zamanlı bir işte çalışmayı düşünürdüm.
  16. Çocuğum olsaydı, oğluma nasıl davranıyorsam kızıma da aynı şekilde davranırdım.
-

## Scientific article review platform using generative artificial intelligence to streamline the peer review process

German Cuaya-Simbro <sup>1\*</sup>, Serguei Drago Domínguez Ruíz <sup>1</sup>

<sup>1</sup>Tecnológico Nacional de México ITS del Oriente del Estado de Hidalgo, Ingeniería en Sistemas Computacionales, Hidalgo, México

### ARTICLE HISTORY

Received: Nov. 22, 2024

Accepted: Aug. 1, 2025

### Keywords:

Web platform,  
Intelligent agent,  
Automatic evaluation,  
Improving objectivity.

**Abstract:** This study introduces a novel Generative Artificial Intelligence (GAI) platform designed to streamline the peer review process. By analyzing a case study of 10 scientific articles, we demonstrate that GAI effectively evaluates article quality and pinpoints specific areas requiring improvement. Our platform achieves an average similarity of 63.6% with human reviewers, enabling the automation of routine evaluation tasks while enhancing both efficiency and objectivity. By drawing on recent generative AI benchmarks across research support, educational assessments, reviewer matching, and large-scale application studies, we demonstrate a focused, practically validated solution that not only aligns with but slightly outperforms general GAI performance levels, offering a transformative approach to real-world manuscript evaluation.

## 1. INTRODUCTION

Double-blind peer review is essential for maintaining the quality and advancing the knowledge presented in scientific articles. However, this process often faces challenges related to review speed, which is contingent upon the availability of reviewers and their time commitments. Additionally, conflicts of interest may arise, potentially undermining the objectivity of the review process. To address these issues, we propose leveraging technologies such as artificial intelligence. This research demonstrates how to integrate and leverage current technologies like Generative Artificial Intelligence (GAI) to develop an agent that mimics the role of a scientific manuscript reviewer, thereby streamlining the review process. The virtual agent accelerates the review process by ensuring adequate reviewer coverage, and to test and assess the virtual agent's effectiveness, we developed a specialized scientific manuscript review platform. Our findings demonstrate the potential of custom-built platforms to rapidly integrate GAI and the feasibility of using this technology to enhance collaborative processes. Finally, we also analyzed the similarity measure to look for patterns of characters within a text, which is of interest to find not only exact matches between two texts, but also to have a measure of approximation between them when the match is not perfect.

Artificial Intelligence (AI) has proven to be fundamental for the automation of different processes in the industry, as discussed in Jan *et al.* (2023). Their research highlights how

---

\*CONTACT: German Cuaya-Simbro ✉ [gcuaya@itesa.edu.mx](mailto:gcuaya@itesa.edu.mx) 📍 Tecnológico Nacional de México ITS del Oriente del Estado de Hidalgo, Ingeniería en Sistemas Computacionales, Hidalgo, México

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

advancements in AI, machine learning, and data analysis have driven the adoption of Industry 4.0 concepts across numerous sectors, enabling waste-free and timely production. Other studies, such as Saranya & Subhashini, (2023), have highlighted AI's capacity to mimic human intelligence and solve complex real-world problems. While AI models can generate outputs without human intervention, their complexity can hinder understanding. Given its demonstrated success in various domains, AI presents a promising avenue for further exploration, particularly in the aforementioned areas.

Generative Artificial Intelligence (GAI), an evolution of AI, has emerged as a powerful tool with the potential to revolutionize various domains. GAI's ability to generate creative and original content has led to numerous innovative applications across various fields. In the decision support for research activities, recent studies have contributed complementary insights into how both generative AI models and structured decision-support platforms can enhance evidence synthesis and policy analysis. Hossain (2024) explored the integration of ChatGPT and related generative AI tools within systematic review workflows, highlighting their capacity to accelerate literature screening and data extraction while cautioning against methodological pitfalls and ethical concerns. Cohen & Moher (2025) examine the broader implications of generative AI in academic writing, underscoring the need for rigorous human oversight and transparent reporting to mitigate the risks of plagiarism, hallucinations, and inaccurate citations. Important highlights are reported in “The Ethical Implications of Using AI in Qualitative Research,” 2025, which critically examines the ethical implications of deploying AI—particularly generative language models—throughout qualitative research processes, emphasizing the necessity of transparency, informed consent, and human oversight to preserve validity and trust. These works map a trajectory from GAI-driven synthesis to ethical guardrails, informing strategies for high-quality decision making in research contexts. Building on this foundation, our manuscript proposes the development of an integrated generative AI system specifically designed to automate and augment the peer-review process for scientific manuscripts.

Some specific applications of GAI that align closely with our proposal include the study of Joe Deavany & Grossfeld (2023), which positions the generative AI as a “copilot” for strategic intelligence analysts, demonstrating that, despite occasional bias or “hallucinations,” AI can automate routine data-gathering so experts focus on high-value, context-driven judgments. On the other hand, Morande (2023) evaluates multiple generative-AI models across key research-support tasks, such as drafting literature reviews, generating hypotheses, and summarizing findings, and provides practical guidelines for integrating these tools responsibly into academic workflows. Fischer *et al.* (2024) examine how AI-generated student submissions align with instructor ratings, identifying both strengths and limitations of automated formative evaluation. Checco *et al.* (2021) investigate the precision and ethical implications of semi-automated systems for matching manuscripts to reviewers, highlighting scalability challenges when handling thousands of submissions. Sengar *et al.* (2024) synthesize findings from over 1,300 studies on generative AI applications, mapping out prevailing performance trends, common pitfalls like model “hallucinations,” and opportunities for domain-specific adaptation.

In contrast, our platform brings these insights together by embedding a generative-AI agent directly into the peer-review workflow, achieving a 63.6% average agreement with human reviewers. This approach delivers a targeted, reproducible solution that not only meets these established benchmarks but also advances consistency, transparency, and objectivity in scientific manuscript evaluation.

To validate the efficiency of using GAI in the review process, we developed a scientific article review platform, as commercial systems such as Open Journal Systems (Open Journal Systems, n.d.), Editorial Manager (Aries Systems, n.d.), and ScholarOne Manuscripts (Silverchair Support, n.d.), often lack the flexibility to integrate algorithms, custom AI, or natural language



processing tools like GAI. The custom web platform developed in this research enabled us to leverage GAI through a virtual agent to automate the review of scientific texts. The agent was able to detect grammatical errors, inconsistencies, and provide a critical review of the text, effectively emulating the work of a journal reviewer. This advancement in text analysis not only accelerates the review process but also empowers journal editors with an additional tool to enhance the quality and objectivity of their reviews.

## 2. METHOD

The following section outlines our research methodology, detailing the design, sample selection, data collection procedures, and evaluation workflow employed to develop and assess AgentRevIAG.

### 2.1. Study Design

The primary objective of this study was to develop and evaluate a generative AI agent, AgentRevIAG, and develop a web platform to integrate the agent to assist and automate the scientific peer review process. To evaluate the closeness between the descriptions of the human experts and the responses of the AI agent, we get a quantitative measure, the BERT similarity measure. And we architected the platform using a MERN (MongoDB, Express.js, React, Node.js) stack with microservices for modularity, to allow embedding generative AI directly into editorial workflows. In the following, each aspect relevant to the creation of the Web Platform and the agent performance evaluation measure is described in more detail.

#### 2.1.1. Web platform

Web platforms are digital environments that offer services, tools, and resources to users through the Internet. These platforms facilitate interaction, collaboration, and transactions between individuals, businesses, and organizations in various domains, such as social media, e-commerce, and online education.

#### 2.1.2. Frameworks

Frameworks are essential tools in web development that streamline the process of building applications by providing predefined structures and common functions. This research utilizes Next.js and Express.js, two popular frameworks. Next.js is a React framework that combines advanced features like Server-Side Rendering (SSR) and Client-Side Rendering (CSR) to deliver smooth user experience and optimize search engines. Express.js is a Node.js framework that simplifies the creation of web applications and APIs, focusing on handling requests and responses and managing routes and middleware.

#### 2.1.3. Generative Artificial Intelligence

Generative Artificial Intelligence (GAI) focuses on creating models and systems capable of generating new and creative content, such as images, music, text, and more. GAI relies on advanced algorithms and machine learning techniques to mimic and emulate human creativity by identifying patterns and features in training data (Gozalo-Brizuela & Garrido-Merchán, 2024).

#### 2.1.4. GAI tools

This study integrates several leading generative AI tools to power our manuscript-review platform. Chat PDF (ChatPDF GmbH, n.d.) streamlines PDF comprehension by automatically extracting and synthesizing key concepts. Sharly AI (Sharly, n.d.) processes large document sets with advanced machine-learning algorithms to deliver tailored, context-rich insights. Gemini (Google, n.d.) successor to LaMDA and PaLM, offers a multimodal language model rivaling GPT-4. ChatGPT (OpenAI, n.d.) excels at natural-language understanding and generation, enabling dynamic, personalized dialogue across simple queries to complex discussions. While several GAI tools are available, many require direct use on their respective

platforms. By utilizing an API to access a GAI tool, our platform offers flexibility and convenience.

### 2.1.5. Similarity measure BERT

To evaluate the closeness between the descriptions of the human experts and the responses of the AI agent, we get a quantitative measure. Conventional techniques for assessing sentence similarity frequently struggle to grasp the intricate nuances and semantic connections found within sentences. With the rise of Transformer-based models such as BERT, there is potential to improve sentence similarity measurements with increased accuracy and contextual awareness. In transformer-based sentence similarity, two input sentences are encoded into fixed-size representations, and their similarity is then measured.

Then we describe the general approach using a pre-trained transformer model, BERT:

1. Preprocess Input Sentences: Tokenize the input sentences into tokens. Add special tokens at the beginning and the end of each sentence. Pad or truncate the token sequences to a fixed length.
2. Encode Sentences: Pass the tokenized sentences through the pre-trained transformer model BERT to obtain contextual embeddings for each token.
3. Calculate Similarity: Measure the similarity between the two sentence embeddings using a similarity metric like cosine similarity or Euclidean distance.

In our case, we used the cosine similarity function of the Sklearn Metrics Pairwise module. Cosine similarity is particularly useful in this context because it compares the similarity between two feature vectors in a multidimensional space, focusing on orientation rather than magnitude. Equation 1 represents how to compute this measure.

$$\text{similarity measure} = \frac{A \cdot B}{\|A\| \|B\|} \quad (1)$$

where  $A \cdot B$  represents the dot product between vectors  $A$  and  $B$ , and  $\|A\|$  and  $\|B\|$  are the norms of vectors  $A$  and  $B$ , respectively.

## 2.2. Sample Characteristics

Due to the journal's strict confidentiality policy, from which we get the manuscripts analyzed in this study, we are unable to include or discuss manuscripts beyond the original set of 10. To ensure transparency within these constraints, we present Table 1, which summarizes key characteristics of each manuscript-topic area, and document type-thereby contextualizing our sample without breaching confidentiality.

**Table 1.** Manuscripts' characteristics.

	Topic area	Document type
Manuscript 1	Trends, Technologies, and Automation in Industry and Industry 4.0	Original Research
Manuscript 2	Trends, Technologies, and Automation in Industry and Industry 4.0	Original Research
Manuscript 3	Trends in Business Administration and Management	Original Research
Manuscript 4	Trends in the food industry in the new normal	Theoretical
Manuscript 5	Trends in Business Administration and Management	Systematic Reviews
Manuscript 6	Trends in Business Administration and Management	Original Research
Manuscript 7	Trends in the food industry in the new normal	Original Research
Manuscript 8	Trends, Technologies, and Automation in Industry and Industry 4.0	Original Research
Manuscript 9	Trends, Technologies, and Automation in Industry and Industry 4.0	Original Research
Manuscript 10	ICTs applied to Tourism Services	Narrative Reviews

## 2.3. Data Collection

Human reviewers evaluated each manuscript using a standardized form covering 10 criteria (scores and comments), each was rated on a scale from 1 (poor) to 5 (excellent), with space provided for comments to justify each rating, the criteria are: Quality of the article's abstract, Contribution to the journal's scope, Description of the methodology, Scientific rigor of the article, Support and evidence provided in the article, Article structure, Writing style and clarity, Literature review, Overall evaluation of the article, and Reviewer's confidence and expertise.

Manuscript PDFs were uploaded to the web platform developed, where we extracted metadata and content for processing. Posteriorly, AgentRevIAG was invoked with structured prompts to generate AI-based evaluations. All prompt texts and API integration details are provided in the next sections.

## 2.4. Evaluation Workflow

AgentRevIAG analyzes articles page by page, providing an overall assessment of their quality. It assigns an objective rating based on criteria such as:

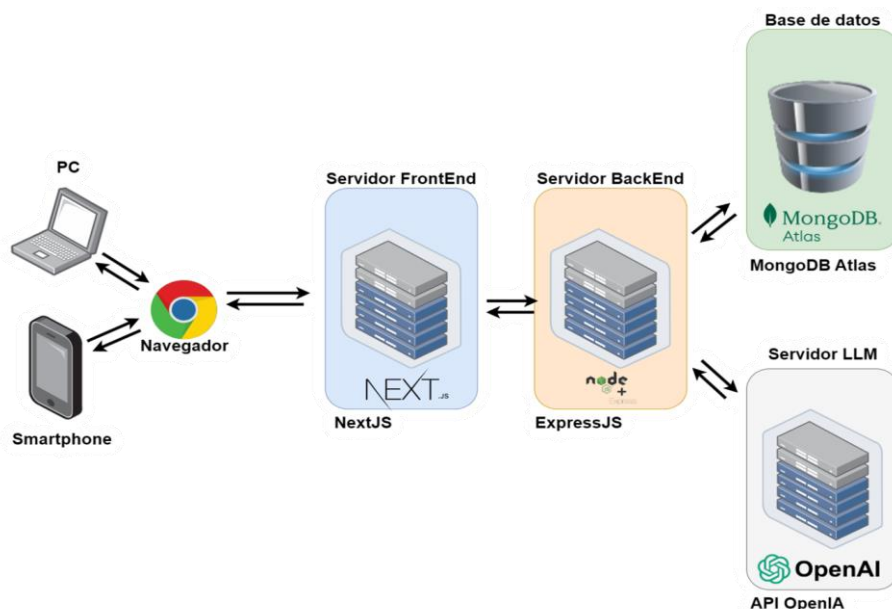
- Structure: logical flow, section completeness, format adherence.
- Clarity: readability, coherence, grammar accuracy.
- Originality: novelty relative to existing literature.
- Rigor: methodological soundness and statistical validity.

AgentRevIAG also offers detailed feedback on specific areas that need improvement, including suggestions for structure, coherence, clarity, and other aspects.

## 3. RESULTS

### 3.1. Article review platform

We developed the scientific article review system using the MERN STACK architecture, which consists of MongoDB, Express.js, React, and Node.js. This architecture, combined with microservices, provides a robust foundation for the system (Figure 1).



**Figure 1.** Platform's microservices architecture.

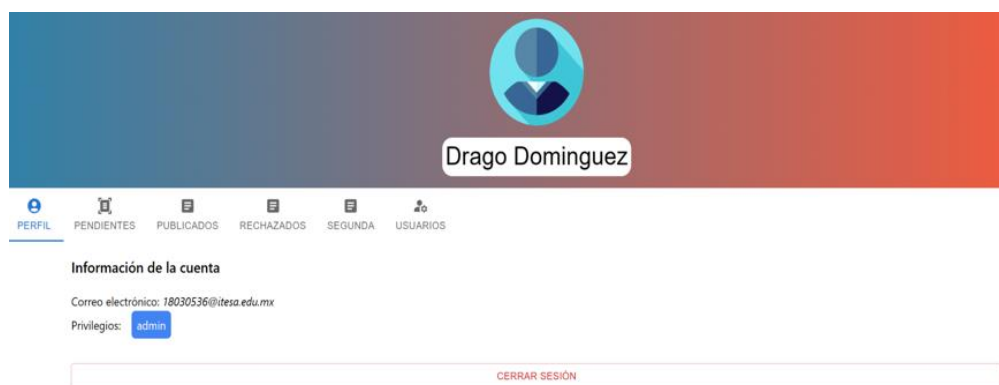
We implemented several basic functionalities in the web platform, including comprehensive article lifecycle management, PDF file upload and storage, a peer review system, and an intuitive user interface. Three user roles were established: administrator, reviewer, and author.

Each role has specific objectives, and only the administrator can access the virtual evaluation agent powered by GAI, referred to as AgentRevIAG.

- 1) The 'Author' page is a dedicated space on the web platform where researchers can view their profile information and upload a new manuscript.
- 2) The 'Reviewer' interface allows reviewers to assess the scientific quality of an article and provide a rating and comments, using the form commented in section 2.3.
- 3) The 'Administrator' role corresponds to the journal editor, who is the only user with access to AgentRevIAG.

The administrator dashboard, shown in [Figure 2](#), allows the editor:

- a) Get article information: Relevant information about the article is displayed, such as its abstract and the date of creation/update. The article category is also included.
- b) Get author information: The name of the author of the article is displayed in a box next to the category.
- c) View article status: An indicator is displayed that reflects the status: under review, to be assigned, rejected, or accepted.
- d) Assign an article to a reviewer: Allows the administrator to select a reviewer to assign the article for review.
- e) Generate an automatic evaluation with AgentRevIAG: Calls the agent to review a manuscript.
- f) Make comments on an article: Allows the re-viewer to send comments on an article to the author.



**Figure 2.** Administrator role main interface.

### 3.2. Development and integration of AgentRevIAG

We developed AgentRevIAG, a GAI-powered review agent, by providing it with structured prompts to generate relevant responses. To integrate AgentRevIAG into the platform, we utilized the OpenAI ChatGPT API.

#### 3.2.1. Configuration of the AgentRevIAG review agent

The following process was followed to develop AgentRevIAG:

1. We designed conversation prompts to guide ChatGPT's interactions. These prompts established ChatGPT's role as a scientific article review expert and elicited opinions on various aspects, including content, methodology, and originality.

**Prompt:** "Assume the persona of a seasoned peer reviewer for an international scientific journal."

2. To evaluate specific topics, we adapted AgentRevIAG by designing additional conversation prompts. These prompts focused on asking clear and direct questions to determine the relevant area of the summary.

**Prompt:** "I need assistance reviewing the following scientific article abstract. From the list below, identify which field it belongs to—Civil Engineering, Food Science, Business Administration, Logistics, Tourism, Industry 4.0, Educational Research, Basic Sciences, Computer Systems, Mechatronics, Electromechanics, or Business Management. Respond with only the field name."

3. AgentRevIAG assessed articles' quality, validity, and relevance by evaluating methodology, originality, results, conclusions, presentation, ethics, and compliance with publication guidelines.

**Prompt:** "I need a detailed critique of this scientific article (I will send it page by page). For each page, decide whether it merits publication, assign it a score out of 100, and justify your decision. Concentrate on errors only—unnecessary or distracting text, overly long or unclear sentences, and any potential plagiarism. Respond solely with the errors you identify."

4. We configured AgentRevIAG to provide an overall assessment, including a quantitative rating and justification.

**Prompt:** "Give your overall assessment of the article by answering just these questions:

Does the article deserve publication?

What score do you assign (X/100)?

Why have you given this score?"

The above leads to the construction of a prompt that is passed to the GAI so that AgentRevIAG can perform a general and objective evaluation of a scientific article.

### 3.2.2. Interaction with OpenAI

We integrated the web platform with ChatGPT using the OpenAI API. Each interaction was structured as a conversation, and the model's responses were captured for analysis. To ensure API compliance, we implemented a mechanism to limit the number of tokens in requests, trimming characters from the end of the text if necessary.

### 3.2.3. Use of the AgentRevIAG

Figure 3 shows the interface for using AgentRevIAG, the GAI-powered review agent. AgentRevIAG analyzes articles page by page, providing an overall assessment of their quality. It assigns an objective rating based on criteria such as Structure, Clarity, Originality, and Rigor, which is in accordance with the evaluation form used by the human reviewers.



Figure 3. Interface for using AgentRevIAG.

## 3.3. Evaluation of AgentRevIAG

To evaluate AgentRevIAG's effectiveness, we conducted a case study analysis of 10 scientific articles. We compared AgentRevIAG's evaluations to those provided by reviewers at a Mexican



scientific journal. This comparison helped us assess GAI's ability to analyze and rate articles, identify strengths and limitations, and identify areas for improvement.

**Table 2** presents 3 random examples of the manuscript reviewed by two persons, and presents a comparison of AgentRevIAG's review, and the last column presents a general qualitative discussion about AgentRevIAG's performance.

**Table 2.** *Examples of qualitative analysis.*

Evaluation Reviewer 1	Evaluation Reviewer 2	Evaluation AgentRevIAG	Qualitative Analysis
<b>Manuscript 1</b>			
I recommend acceptance of the work for REISITAL, perhaps with some modifications about the observations made. The work has sufficient theoretical and methodological support to be published.  Comment for (Review of bibliography): Place the references in APA 7th Edition format for the books used, regarding the author Carreras it is suggested to handle as a parenthetical short quotation, therefore, it should be integrated at the end of the text found between quotation marks, regarding the reference Zichermann and Cunningham 2011, review the quotation, due to the fact that the surnames are inverted, review the year for the reference and quotation of the author Gallejos.	Review type in summary <design>, it is suggested to use a justified text in the text (line on the left is observed) It is suggested to pay attention to the comments previously indicated.	According to the set of inconsistencies mentioned above, it is essential that the author makes the necessary corrections before considering publication of the article. There are significant problems, especially with respect to clarity in the introduction, abstract and conclusions. In addition, inconsistencies in grammar and spelling errors, as well as overuse of keywords, also weaken the overall quality of the paper.  That said, as it stands, the grade I would give the paper would be 60/100. This rating is justified based on the numerous problems identified that need to be addressed before the article is ready for publication.  Finally, despite the problems noted, the topic of the article is relevant and has the potential to contribute significantly to the field of study. Therefore, with appropriate corrections and suggested improvements, I am confident that the article would have a great impact and expand the body of knowledge in the area of study.	AgentRevIAG provides a more detailed and critical evaluation, identifying specific areas for improvement. While human reviewers highlight positive aspects, AgentRevIAG's evaluation can serve as a guide for overall quality. Although the scores differ slightly, both human reviewers and AgentRevIAG agree on the article's acceptance. This suggests that AgentRevIAG's evaluations can align with human judgment.
<b>Manuscript 2</b>			
It is recommended to respect the REISITAL format (remove the title page, acknowledgments and the number of words in the abstract).	Very good work, aligned with the thematic axes of the journal and pertinent from the point of view of the use of technology in teaching.	Based on the problems presented, I believe that the article needs to be thoroughly restructured and revised before it can be considered for publication. The errors and problems noted suggest that it is not up to the level required for publication in its current state.  The rating I would give this article is 60/100. This rating is due to the lack of clarity and consistency, as well as the need for intensive revision to correct errors and to ensure effective delivery of the message and research findings.  In addition, it is important to keep in mind that scientific research needs to be communicated clearly and effectively so that it can be useful and accessible to others in the field, as well as to the general public. The lack of clarity and consistency in the abstract, the excessive use of jargon, and the lack of a clear results section	In this article, there is a more notable difference in what the IAG comments in relation to the REISTAL reviewer, they only agree on what is related to the wording, but considering the acceptance note that was discussed in the example of article 1, both the IAG and the reviewer would accept the article having given 27/45 and 33/45 respectively.

Evaluation Reviewer 1	Evaluation Reviewer 2	Evaluation AgentRevIAG	Qualitative Analysis
		indicate that there is room for improvement in the communication in this article. Therefore, I would recommend that substantial revisions be made to correct these problems prior to publication.	
<b>Manuscript 3</b>			
Very good work, aligned with the thematic axes of the Congress and pertinent from the point of view of the use of technology in teaching.	Revise the wording of bibliographic references and citations according to APA format 7 edition. Not all references are cited. Avoid repeating references (e.g. lines 423 and 435).	Even with the observations, it is evident that there is valuable research work in the article presented, but significant problems in its execution and presentation prevent it from being published in its present form. However, with the aforementioned corrections and modifications, it can be a valuable contribution to the scientific literature.  Rating: 60/100. This rating is due to the number of formatting and structural errors, as well as deficiencies in the presentation and description of methods and results. Readability problems and lack of key sections diminish the overall quality of the manuscript.  The potential is there, but work is needed to polish this study before publication. Significant revisions are needed to meet the standards required for publication.	We observed that two reviewers lacked expertise in the subject area, resulting in superficial evaluations. In contrast, the reviewer with a higher level of expertise provided more accurate assessments that aligned with AgentRevIAG's evaluation. Despite the differing levels of expertise, all reviewers, including AgentRevIAG, agreed on the article's acceptance.

In a general way, after reviewing all GAI agent reviews, we can see that while AI-powered peer review offers significant potential benefits, it is not a replacement for human expertise. The optimal approach is to leverage AI as a tool to augment human capabilities, thereby improving the overall quality and efficiency of the peer review process.

Finally, [Table 3](#) presents a summary of similarity measures, quantitative analysis from all manuscripts reviewed.

**Table 3.** *Quantitative analysis.*

	Similarity measure R1 – R2	Similarity measure R1 - GAI	Similarity measure R2 - GAI	Average similarity of the IAG and both reviewers
Manuscript 1	58.1	71.6	61.2	66.4
Manuscript 2	64.3	59.5	60.6	60.1
Manuscript 3	50.2	73.8	65.3	69.6
Manuscript 4	66.0	63.1	60.5	61.8
Manuscript 5	64.1	59.2	61.8	60.5
Manuscript 6	48.4	60.5	42.6	51.6
Manuscript 7	44.1	59.7	68.5	64.1
Manuscript 8	67.3	69.4	64.2	66.8
Manuscript 9	70.1	75.9	79.5	77.7
Manuscript 10	45.3	64.1	51.6	57.8
Average similarity measure R1 – R2	57.8		Average similarity of the IAG and both reviewers	63.6
Standard deviation	9.8		Standard deviation	7.1

According to the results presented in Table 3, across the four topic areas, our generative-AI agent not only matches but, in many cases, exceeds human reviewer consistency. For instance, manuscripts on “Trends, Technologies, and Automation in Industry and Industry 4.0” achieved the highest inter-human agreement (Reviewer 1 vs. Reviewer 2: ~65%) and an even stronger AI-to-human similarity (~67.8%). In contrast, articles on “ICTs applied to Tourism Services” showed the lowest human agreement (~45.3%) and a correspondingly lower AI-to-human similarity (~57.8%), suggesting that less structured or more qualitative content yields higher variability overall. When we break down by document type, “Original Research” papers saw the highest AI-human alignment (~65.2%), while “Narrative Reviews” were the most challenging (~57.8%), reflecting their broader interpretive scope. It is worth mentioning, across every subgroup, the AI-agent’s average similarity with both reviewers (63.6% overall) consistently meets or exceeds the human reviewers’ agreement levels, underscoring the platform’s robustness. These patterns suggest that our GAI tool performs particularly well on tightly structured, data-driven manuscripts, and though it remains slightly less consistent on open-ended or thematic reviews, it still provides acceptable alignment—demonstrating its potential as a reliable assistant in the peer-review process.

#### 4. DISCUSSION and CONCLUSION

This work presents AgentRevIAG, a generative-AI review agent seamlessly integrated into a custom web platform for scientific manuscript evaluation. Our case study shows that AgentRevIAG’s assessments align closely with human reviewers—achieving a 63.6% average similarity—and deliver consistent, objective feedback that can augment editorial decision-making.

Our evaluation of AgentRevIAG across 10 manuscripts demonstrates that it performs matching or slightly exceeding performance baselines established by Morande (2023), Fischer *et al.* (2024), Checco *et al.* (2021), and Sengar *et al.* (2024), which average around 59-62% accuracy in diverse GAI tasks. This consistency suggests that embedding GAI directly in the peer review workflow produces reliably consistent assessments. Importantly, while human reviewers showed more variability on qualitative or narrative reviews (e.g., Tourism Services), AgentRevIAG maintained comparable alignment, suggesting robustness even on less-structured content.

Ethical considerations, such as potential bias or “hallucinations” in training data, remain an open concern, referring to warnings from Cohen & Moher (2025), and the ethical review presented in (Hitch *et al.*, 2025). Moreover, confidentiality constraints limited our sample to ten manuscripts, constraining the generalizability of results, which we consider a relevant limitation of our research.

Despite these limitations, the tool offers clear advantages. First, it automates routine evaluation tasks, speeds review turnaround, and provides objective, reproducible feedback. By situating our findings within the broader literature on AI-assisted review and automated evaluation processes, we demonstrate both the feasibility and the practical value of a domain-specific GAI review assistant.

Future work should focus on expanding to larger and multilingual datasets, incorporating explainability mechanisms to surface the agent’s reasoning, conducting editor and author user-experience studies, and exploring next-generation GAI models (e.g., ChatGPT v4.0) to further enhance accuracy and applicability.

#### Acknowledgments

This work has been carried out thanks to funding granted by the “Convocatoria Proyectos de Investigación Científica, Desarrollo Tecnológico e Innovación 2024” for the project “Desarrollo de una plataforma de revisión de artículos científicos incorporando Inteligencia Artificial Generativa (IAG)” with id: 19506.24-PD.

We also acknowledge the joint effort of all the authors of the manuscript, who belong to the Computer Systems Engineering Department of the Tecnológico Nacional de México / ITS del Oriente del Estado de Hidalgo.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

### Contribution of Authors

**German Cuaya-Simbro:** Investigation, Supervision, Resources, Formal analysis, and Writing-original draft. **Serguei Drago Domínguez Ruíz:** Methodology, Visualization, Software development, and Validation.

### Orcid

German Cuaya-Simbro  <https://orcid.org/0000-0001-6303-154X>

Serguei Drago Domínguez Ruíz  <https://orcid.org/0009-0001-9245-6704>

### REFERENCES

- Aries Systems. (n.d.). *Editorial Manager* [Cloud-based manuscript submission and peer review system]. Aries Systems Corporation. <https://www.ariessys.com/solutions/editorial-manager/>
- ChatPDF GmbH. (n.d.). *ChatPDF* [AI-powered app]. <https://www.chatpdf.com/>
- Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8, Article 25. <https://doi.org/10.1057/s41599-020-00703-8>
- Cohen, J.F., & Moher, D. (2025). Generative artificial intelligence and academic writing: Friend or foe? *Journal of Clinical Epidemiology*, 179, Article 111646. <https://doi.org/10.1016/j.jclinepi.2024.111646>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1810.04805>
- Fischer, I., Sweeney, S., Lucas, M., & Gupta, N. (2024). Making sense of generative AI for assessments: Contrasting student claims and assessor evaluations. *The International Journal of Management Education*, 22(3), Article 101081. <https://doi.org/10.1016/j.ijme.2024.101081>
- Google. (n.d.). *Gemini* [Large language model]. Google AI. <https://gemini.google.com>
- Gozalo-Brizuela, R., & Merchan, E.E.G. (2024). A Survey of Generative AI Applications. *Journal of Computer Science*, 20(8), 801-818. <https://doi.org/10.3844/jcssp.2024.801.818>
- Hitch, D., Richards, K., Gupta, A., & Thanekar, U. (2025). The ethical implications of using AI in qualitative research. In A. Gupta (Ed.), *Artificial intelligence (AI) in social research* (pp. 137-149). CABI Publishing. <https://doi.org/10.1079/9781800626607.0013>
- Hossain, M.M. (2024). Using ChatGPT and other forms of generative AI in systematic reviews: Challenges and opportunities. *Journal of Medical Imaging and Radiation Sciences*, 55(1), 11-12. <https://doi.org/10.1016/j.jmir.2023.11.005>
- Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., & Kuusk, A. (2023). Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert Systems with Applications*, 216, Article 119456. <https://doi.org/10.1016/j.eswa.2022.119456>
- Joe Devanny, H.D., & Grossfeld, E. (2023). Generative AI and intelligence assessment. *The RUSI Journal*, 168(7), 16-25. <https://doi.org/10.1080/03071847.2023.2286775>
- Morande, S. (2023). Benchmarking Generative AI: A comparative evaluation and practical

- guidelines for responsible integration into academic research. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4571867>
- Open Journal Systems. (n.d.). OJS (Version 3.1) [Open source publishing platform]. <https://openjournalsystems.com/>
- Saranya, A., & Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7, Article 100230. <https://doi.org/10.1016/j.dajour.2023.100230>
- Sengar, S.S., Hasan, A.B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: A systematic review and applications. *Multimedia Tools and Applications*, 84, 23661-23700. <https://doi.org/10.1007/s11042-024-20016-1>
- Sharly. (n.d.). *Sharly AI* [AI tool]. VOX AI Inc. <https://sharly.ai/>
- Silverchair Support. (n.d.). *ScholarOne Manuscripts* [Workflow management system]. Silverchair. <https://www.silverchair.com/products/scholarone-manuscripts/>



## Developing Turkish-Sign-Language Self-Efficacy Scale for Learners based on various test theories

Pelin Piştav Akmeşe<sup>1</sup>, Asiye Şengül Avşar<sup>2\*</sup>, Nilay Kayhan<sup>3</sup>,  
Necla Işıkdöğün Uğurlu<sup>4</sup>, Ayşen Zeynep Oral<sup>5</sup>

<sup>1</sup>University of Ege, Faculty of Health Sciences, Department of Audiology, İzmir/Türkiye

<sup>2</sup>Recep Tayyip Erdogan University, Faculty of Education, Department of Measurement and Evaluation in Educational Sciences, Rize/Türkiye

<sup>3</sup>University of Ankara Yıldırım Beyazıt, Faculty of Health Sciences, Department of Speech and Language Therapy, Ankara/Türkiye

<sup>4</sup>Zonguldak Bülent Ecevit University, Faculty of Education, Department of Special Education, Zonguldak/Türkiye

<sup>5</sup>Hacettepe University, Faculty of Letters, Department of Translation and Interpreting, Ankara/Türkiye

### ARTICLE HISTORY

Received: Nov. 10, 2024

Accepted: Aug. 16, 2025

### Keywords:

Turkish-Sign-Language,  
Self-efficacy,  
Scale development,  
Mokken scale analysis,  
Principal axis factoring.

**Abstract:** Turkish-Sign-Language (TSL) is a natural visuospatial language that the Deaf and hard of hearing use to communicate both with each other and with hearing people. It is important to determine the self-efficacy of individuals learning TSL in order to enhance their effective use of the TSL when their learning process. In this study, we aimed to develop a scale that measures TSL self-efficacy for TSL learners and provide valid and reliable results with different test theories. The study was designed in a quantitative research design, and the participants consisted of 430 university students who were identified through a purposive sampling technique. Automated item selection procedure in the context of Mokken Homogeneity Model-one of the nonparametric item response theory models, and principal axis factoring, convergent and discriminant validity in the context of classical test theory were investigated to present evidence for construct validity. According to the analysis, a measurement tool consisting of 22 items and a three-factor structure was reached. The reliability coefficients of the scores were investigated via Cronbach's Alpha ( $\alpha$ ), Guttman's lambda 2 ( $\lambda$ ), latent class reliability coefficient (LCRC), composite reliability coefficients (CR), and McDonald's omega ( $\omega$ ). The reliability values obtained from these coefficients indicated that the scores obtained from the scale were reliable. As a result, the TSL-Self-Efficacy Scale for TSL learners, which provided valid and reliable results, was successfully developed.

## 1. INTRODUCTION

Communication is a two-way process and involves interaction. Effective communication enables individuals with Deaf and hard of hearing (DHH) to meaningfully participate fully in educational and work environments (National Deaf Center-NDC, 2019). When comparing the communication skills of children with hearing impairment and typically developing children, it

\*CONTACT: Asiye ŞENGÜL AVŞAR ✉ [asiye.sengul@erdogan.edu.tr](mailto:asiye.sengul@erdogan.edu.tr) 📍 Recep Tayyip Erdogan University, Faculty of Education, Department of Measurement and Evaluation in Educational Sciences, Rize/Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

was determined that children with hearing impairment performed lower in communication skills than their non-hearing peers (Tuohimaa *et al.*, 2022). Research reveals that DHH individuals face barriers such as misdiagnosis, marginalization, and discrimination in accessing healthcare services (Wilson-Menzfeld *et al.*, 2025) and communication difficulties with teachers and peers in educational settings (Nikolarazi *et al.*, 2015). Effective communication is based on considering an individual's preferred communication method (Ballenger, 2025).

Hearing impairment encompasses hearing loss from mild to profound levels (Darica & Şipal, 2011). Individuals with hearing impairments need special education and support services. Therefore, they receive education, training, and support services within the scope of the Special Education Services Regulation in institutions and organizations affiliated with the Ministry of National Education (MoNE) in Türkiye (MoNE, 2018). In addition, newborn hearing screening programs are extremely important because early diagnosis of hearing loss and effective intervention programs yield positive results (Marriage *et al.*, 2017). Early intervention services for children with hearing loss should be family-centered; professional help should be provided to support communication and language skills, taking into account the socio-cultural characteristics of families, and inequalities in access to services should be eliminated (Humphries *et al.*, 2022; Murray *et al.*, 2019; Szarkowski *et al.*, 2024). The mother tongue of babies born with severe and profound hearing loss is sign language (Kubuş *et al.*, 2016; Piştav Akmeşe, 2019). Sign languages are natural languages with which DHH people can communicate with each other and with other people (Eryiğit, 2017). It is crucial for children to use sign language in natural environments, such as at home and in daily life. Media and television should ensure barrier-free access by including sign language translation and detailed subtitles in health, education, and entertainment programs. Nearly all DHH children who lack early intervention and auditory-verbal education, and who attend schools for DHH students, rely on sign language for communication (Gürboğa & Kargın, 2003).

Sign language is both an academic tool and a mother tongue for DHH children, facilitating communication and participation in family and community life (Alfano *et al.*, 2022; Musyoka, 2022). It is stated that inclusive education practices for DHH students need to be improved and in this context, sign language practices should be included within the scope of staff competencies (Beal *et al.*, 2024; Jones *et al.*, 1997; Salehomoum, 2020). In an inclusive educational environment, DHH students should be provided with equal access to the curriculum as their hearing peers (Tang, 2024). To effectively teach subjects to DHH individuals, teachers need to recognize the unique learning needs of their students and use specific strategies (Bintoro *et al.*, 2023). Therefore, qualified inclusive teacher training programs are necessary to provide a good education to students with hearing impairments (Guardino, 2015).

Turkish-Sign-Language (TSL) is "a natural, visual-spatial language that the Deaf community in Türkiye uses, and that reflects the culture of the Turkish Deaf community with its distinctive grammar, vocabulary and usage features" (Kubuş *et al.*, 2016). TSL is the mother tongue of DHH individuals of Turkish citizens and can be expressed as an important language that should be known and accepted in both educational environments and health social services and society in general, in the development of literacy and numeracy competencies. Until 2016, Turkish was the official and predominant language of education in schools for DHH individuals in Türkiye, and only the auditory-verbal method was used. Therefore, children who learnt sign language in their schools only for communication purposes experienced serious drawbacks in academic learning (Dikyuva *et al.*, 2015; Piştav Akmeşe, 2019). Furthermore, sign language courses were introduced as elective courses in higher education institutions in the 2013-2014 academic year in Türkiye. Since 2016-2017, they have been a compulsory 2-credit course in the 3rd semester of special education teaching programs under the name "Turkish-Sign-Language" (Council of Higher Education, 2018). Teachers who graduate from special education teaching undergraduate programs by taking this course are expected to meet the learning outcomes in the TSL course content.

The limited use of sign language in social life creates barriers for DHH individuals, hindering their participation. In different studies conducted by Hammer (1998) and Ubido Friends (2002), patients reported that they had difficulty accessing interpreter services in the field of health. When the school environment is considered for a DHH child, it is an environment where he/she is expected to adapt, understand and be understood. In schools, there should be arrangements to ensure the adaptation and full participation of DHH students, quality educational adaptations for their career planning, curriculum structuring, and, in short, full participation in academic and social environments (Danermark *et al.*, 2001; Kim *et al.*, 2015; Scott & Hoffmeister, 2017). It is important to answer the question of how high self-efficacy of those who have learned sign language and how they competent and successful perceive themselves to be when communicating with a DHH student, patient, child, or adult (McDermid, 2020). Researchers (Dammayer *et al.*, 2018; Fitriyani *et al.*, 2024; Piştav *et al.*, 2018) have shown that hearing individuals who later learn sign language and develop high sign language self-efficacy can increase mutual understanding by communicating effectively with DHH individuals. Both the sign language skills of professionals who interact with hearing-impaired individuals and their self-efficacy in acquiring and applying these skills are of great importance.

Bandura (1997) defined perceived self-efficacy as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (p. 3). According to Baumeister and Vohs (2007), self-efficacy is an individual's ability to achieve what is expected of them through their own actions and to have the expected impact. The effort a person makes, the level of stress they experience, and their ability to deal with a challenging job also vary according to this level of belief. To exemplify this, individuals who have a high level of self-efficacy also have lower anxiety levels (Avşar & Sevim, 2022). There are four important sources of information that influence self-efficacy, including mastery experience, vicarious experience, social persuasion, and physiological and affective states (Bandura, 1997). It has been stated that mastery experiences are the strongest source of self-efficacy (Artino, 2012; Bandura, 1997; Pfitzner-Eden, 2016; Usher & Pajares, 2008). It is expected that individuals with high self-efficacy in any subject will show high performance in that subject.

People who are not hearing impaired may have a high level of self-efficacy in oral communication, while they may have a low perception of self-efficacy in the use of TSL. Identifying individuals with low self-efficacy in TSL will negatively impact their ability to communicate effectively with DHH individuals. To strengthen their TSL self-efficacy, these individuals' self-efficacy should be assessed during TSL training. At this stage, there is a need for a measurement instrument for TSL self-efficacy that aligns with the theoretical framework of self-efficacy.

There is a large body of research (Almayez *et al.*, 2025; Kim, 2024; Ma, 2022; Panjie & Velarde, 2025; Sun & Mu, 2023) reporting that self-efficacy is associated with learning a new language, and some of these studies indicate that individuals with high self-efficacy learn a language more easily. However, self-efficacy research on sign languages is limited. Examples of research on the relationship between sign language and self-efficacy include the Development of the American Sign Language Self-Efficacy (ASL) Scale for Deaf Individuals (Reddy, 2011) and the research conducted by Koch (2023) based on the use of timely and authentic feedback in education to increase students' ASL Self-Efficacy. Reddy (2011) developed a six-item measurement tool based on three dimensions of self-efficacy in his study. In this research, the focus was on the source of self-efficacy, mastery experience, while developing the scale. Koch (2023) pointed out that students must be confident in their skills to gain proficiency in ASL, students must be given timely and effective feedback in their sign language learning, and ASL self-efficacy must be measured in their sign language learning. For TSL, a valid and reliable scale has been developed to measure attitudes toward TSL in Türkiye (Bilgiç *et al.*, 2021), predicting whether teachers and prospective teachers accept and incorporate sign language education. Although measuring attitudes toward TSL is important,

there is no scale that measures the TSL learners' self-efficacy, such as university students, teachers, adults, or healthcare professionals. As adults learning TSL increase their self-efficacy in communicating in TSL, they will be able to communicate more effectively with native sign speakers, whether in their professional lives or in everyday life.

Considering that sign language is undoubtedly a very important life skill for DHH children, it is essential to determine the self-efficacy towards TSL of the individuals who learn TSL. In this way, reformative studies and activities can be organized for individuals who have low self-efficacy for TSL, and thus, qualified TSL education of DHH children can be supported. The TSL self-efficacy of teachers who prepare DHH individuals for the future is important in communicating with DHH individuals. In addition, for healthcare professionals who will work with individuals with DHH due to their profession, determining TSL self-efficacy while taking TSL courses is also important to increase the quality of services to be provided in the future. For these reasons, the aim of this study is to develop the Turkish-Sign-Language Self Efficacy Scale (TSL-SES) for Learners, which yields valid and reliable results using different test theories and is expected to provide consistency and comparability across groups with different characteristics thanks to its theoretical basis.

## 2. METHOD

### 2.1. Research Model and Participants

This research is a correlational design, which is a quantitative approach that explores associations among multiple variables without the researcher influencing the outcomes (Fraenkel *et al.*, 2012). This method was deemed for examining the relations between scale items, latent constructs, and the overall structure of the TSL-SES.

The number of items was taken into account in determining the sample size. In scale development studies, it is generally recommended to reach at least 10 times the number of items (Cohen *et al.*, 2013). Purposive sampling was used in this study, which allows for the selection and in-depth analysis of information-rich situations in line with the study's purpose (Fraenkel *et al.*, 2012). The participants of the research consisted of 487 university students who had taken TSL courses at the Faculty of Education, Faculty of Health Sciences, and Faculty of Dentistry, and who were informed about the study beforehand, participated in the research voluntarily and online. The number of participants whose data were analysed was 430 after excluding participants whose scores were aberrant. Information on the demographic characteristics of the participants is presented in Table 1.

**Table 1.** Demographic characteristics of participants.

University	N	%
Bülent Ecevit University	125	29.07
Ege University	266	61.86
Others	39	9.07
Gender	N	%
Female	285	66.30
Male	145	33.70
Is there a DHH person in your family?	N	%
Yes	16	3.70
No	414	96.30
Did you take sign language training before university?	N	%
Yes	200	46.50
No	230	53.50
Departments	N	%
Nutrition and dietetics	51	11.86

Dentistry	79	18.37
Physical therapy and rehabilitation	21	4.88
Special education training	272	63.26
Others	7	1.63

The highest participation was from Ege University (61.86%) and mostly female (66.30%). Most participants (96.30%) had no DHH family members. Those with (46.50%) and without (53.50%) sign language training before were nearly equal. The special education department had the highest participation (63.26%).

## 2.2. Data Collection Process and Data Collection Tool

Ethical approval for data collection in this study was granted by the Recep Tayyip Erdoğan University Social and Human Sciences Ethics Committee, based on the decision dated June 14, 2022, with reference number 2022-147.

It was aimed to develop a scale that measures self-efficacy for TSL for learners in a valid and reliable way in this research. There was no other suitable scale for this purpose in the literature. Therefore, the concurrent validity could not be investigated. In our study, evidence for construct validity was presented using different test theories, which are Classical Test Theory (CTT) and nonparametric Item Response Theory (NIRT).

### 2.2.1. Development stage of TSL-SES for learners items

Several steps were followed in developing TSL-SES for TSL learners based on DeVellis (2017). First, a detailed literature review was conducted. Specifically, items that might be the indicators of self-efficacy were determined in consequence of detailed research. We examined self-efficacy scales for languages, Bandura's theory, and four basic self-efficacy sources according to Bandura (1997). Because mastery experiences are the most important source of self-efficacy, most of the items written were specifically related to mastery experiences.

Since no scale measures the related construct in Türkiye, measurement tools that were developed abroad were examined. A total of 30 items of draft were written by field expert academicians (two associate professors of special education who had studied sign language). Following scale development guidelines (Crocker & Algina, 1986), the items were reviewed by two field expert academicians (two assistant professors in special education and not item writers). These experts were asked to evaluate whether the written items were suitable for the measured structure. The reliability between the experts' coding was determined with the

$$\frac{\text{Number of agreements}}{\text{Number of agreements} + \text{Number of disagreements}} * 100$$
 formula, and the obtained value was found to be 0.87. Since this value is above 0.80, consistency between expert opinions has been ensured (Miles & Huberman, 1994). In addition, two separate panel meetings were set up, including the item authors, a measurement and evaluation expert was established to evaluate the feedback from the experts (including suggestions for revisions to the items), and to examine the items in more detail. In these panel meetings, each expert evaluated each item individually. The experts were given the opportunity to comment freely on the items without any pressure. This method was recommended by DeVellis (2017). The items were thoroughly examined by experts, considering the criteria of whether the scale was a good indicator of the psychological construct, whether it could measure self-efficacy, whether it could measure a single psychological indicator, and its linguistic clarity, respectively. As a result of the examinations and discussions, it was agreed that the 24 draft items with a five-point Likert-type scale would be involved in the scale. The scale was designed so that participants could indicate their level of agreement with the items from 1 (Strongly disagree) to 5 (Strongly agree) according to the literature (Erickson & Noonan, 2025). As a result of the experts' examination, the distribution of the remaining items into factors was structured as follows: 14 items on mastery experiences, 4 items on vicarious experiences, 3 items on social persuasion, and 3 items on emotional and



physiological state. These items, which were decided to remain in the scale according to expert opinions, were sent to a Turkish language expert (associate professor) for formal and linguistic expression. As suggested preliminary trial (Crocker & Algina, 1986) with 10 university students (average age 19.40, 6 females, 4 males) confirmed the comprehensibility of all 24 draft items.

### 2.3. Data Analysis

Several validity and reliability analyses were carried out to determine the psychometric characteristics of the scores obtained from the TSL-SES for Learners. First, evidence was searched for construct validity by various methods. For this purpose, Mokken Scale Analysis (MSA), Exploratory Factor Analysis (EFA), and convergent and discriminant validity were investigated. Then, group differences and partial metric invariance were determined to support construct validity proofs. The data were analyzed using EFA in the context of CTT, and the Mokken Homogeneity Model (MHM) within the nonparametric IRT framework.

NIRT, one of the IRT models, was used together with Exploratory Factor Analysis (EFA) in the context of CTT in the development of TSL-SES. CTT and IRT are the most widely used testing theories in education and psychology. CTT assumes that individuals' observed scores are the sum of their true scores and error scores, and item properties and test statistics (such as item difficulty, discrimination index, and measurement error) are estimated depending on the group on which the test is administered (Crocker & Algina, 1986; Embretson & Reise, 2000). IRT estimates an individual's ability or latent traits based on their responses to test items (Embretson & Reise, 2000). Parametric IRT models estimate item and ability parameters independently of the group when model-data fit is achieved (Hambleton *et al.*, 1991). However, for relatively small groups and tests with fewer items, NIRT models stand out because they require fewer assumptions. In this study, the MHM, one of the NIRT models that requires fewer assumptions than the parametric IRT models, was used to development of the TSL-SES (Stochl, 2007). CTT and NIRT were used together in the development of TSL-SES, providing a different perspective on scale development.

The data obtained in accordance with the specified purpose were analyzed using RStudio (Version 2023.06.0.421; RStudio Team, 2023). Various R packages were utilized for data analysis. Descriptive statistics, EFA, and parallel analysis were conducted using the psych package (Revelle, 2023), Mokken analysis was performed using the Mokken package (Van der Ark *et al.*, 2023). Aberrant item scores were identified using the PerFit package (Tendeiro, 2023), while evidence of convergent and discriminant validity and partial invariance were assessed using the lavaan package (Rosseel *et al.*, 2023) and the semTools package (Jorgensen *et al.*, 2023).

#### 2.3.1. Mokken scale analysis

MSA can be thought of as a scaling procedure for both dichotomous and polytomous items according to MHM. MHM is a useful model for scaling in short tests and small samples (less than 500 participants) for establishing construct validity (Galindo-Garre *et al.*, 2014; Palm & Strong, 2007). Since the number of participants in our research was less than 500, we used the scaling technique MHM. In scale development studies, determining the factor structure according to different test theories provides strong evidence for establishing construct validity (Chen *et al.*, 2016; Gao *et al.*, 2024; Pretorius & Padmanabhanunni, 2024). This approach was followed to obtain robust evidence regarding construct validity.

For MHM scaling, the monotonicity assumption must be met, evaluated using *crit* values:  $crit < 40$  is suitable,  $40 \leq crit < 80$  is suspicious, and  $crit \geq 80$  is seriously incompatible (Crişan *et al.*, 2022). Item fit is assessed using *H* scalability coefficients, where  $0.30 \leq H < 0.40$  is poor,  $0.40 \leq H < 0.50$  is moderate, and  $H \geq 0.50$  is high (Sijtsma & Molenaar, 2002). Local independence must be met, and this assumption can be examined by controlling the conditional association

method (Van der Ark, 2023). It must be noted that this method lacks a formal test procedure, leaving it to the researcher to decide if violations warrant rejecting the MHM or removing an item (Koopman *et al.*, 2023).

The *crit* values were examined for the monotonicity assumption in this research. It was determined that the *crit* values were less than 40; in other words, the monotonicity assumption was met. Local independence was examined by conditional association with the *check.ca* method. The relevant results are shown the [Supplementary](#) file.

MHM is an exploratory model with its automated item selection procedure (AISP), especially in scale development studies (Sijtsma & Molenaar, 2002). AISP constitutes one-dimensional Mokken scales by separating items that are not scaled according to MHM; in this way, item clusters are created, which provide information about the dimensionality of the scale.

Item clusters and item selections are made according to the determined cutoff (*c*) value (Sijtsma & Molenaar, 2002) in AISP. This value is increased by 0.05 when starting from at least 0.30, and the results obtained are evaluated (Stochl *et al.*, 2012). The point to take into account is that there is no ideal *c* value. AISP can be summarized as a method in which researchers decide the number of factors according to increasing *c* values. There is no precise ideal value for *c* values. It is suggested that outcomes using values ranging from 0.30 to 0.55 (Emons *et al.*, 2012).

### 2.3.2. Exploratory factor analysis

The principal axis factoring (PAF) model is recommended for EFA in scale development studies (Hair *et al.*, 2019). In this research, PAF with polychoric correlation was used for factor extraction. Oblimin rotation was applied due to the expected correlation between factors, as orthogonal rotation is often unrealistic in social sciences (Hair *et al.*, 2019; Şengül Avşar & Barış Pekmezci, 2022).

Before proceeding to PAF, descriptive statistics regarding item scores were investigated. The findings regarding descriptive statistics were presented in the [Supplementary](#) file. It was seen that the lowest item score average was 2.75 and the highest was 4.18, the lowest skewness value was -0.80 and the highest was 0.24, and the lowest kurtosis value was -0.63 and the highest was 1.33. Since the sample size was greater than 300, the standard values of the skewness and kurtosis coefficients were not considered, and since the absolute values of the skewness coefficients were not less than 2 and the absolute values of the kurtosis coefficients were not greater than 7, it was determined that the item scores did not deviate from the normal distribution (Kim, 2013). In addition, by calculating the Mardia's skewness coefficient for multivariate normality, it was determined that the data did not have a multivariate normal distribution ( $\chi^2 = 4158.19; p < 0.05$ ).

The sample size met the requirement of being ten times the number of items (Kline, 2011), and sample adequacy was confirmed by a Kaiser-Meyer Olkin (KMO) value  $>0.80$  and a statistically significant Bartlett's test (Distefano & Hess, 2005; Kline, 2011). It was determined that the KMO values (three different PAF were conducted) were 0.96 and the Bartlett's tests were statistically significant in our research. This result reveals that the sample size is appropriate for PAF.

### 2.3.3. Convergent and discriminant validity

Convergent and discriminant validity evidence should be presented in scale development studies. For convergent validity, the average variance extracted (AVE) is calculated, while for discriminant validity, the square root of AVE values for each factor is compared with the correlation values between factors (Fornell & Larcker, 1981; Hair *et al.*, 2019).

In the calculation of AVE values, the standard factor loading values obtained from confirmatory factor analysis (CFA) were considered. If the square roots of the AVE values are greater than the correlation values between the factors, this indicates that discriminant validity is established

(Fornell & Larcker, 1981; Hair *et al.*, 2019). Heterotrait-Monotraitratio (HTMT) of the correlations can also be calculated for discriminant validity. A calculated correlation value of less than 0.85 indicates that discriminant validity is achieved (Henseler *et al.*, 2015).

### 2.3.4. Investigation of group differences

One method for establishing construct validity is examining group differences (Cohen *et al.*, 2013). However, it is important to note that evidence from this method alone is insufficient for construct validity and should be supported by additional validation methods (Anastasia, 1976). In this study, the scores of those who had previously received sign language training (n=200) and those who had not received sign language training (n=230) were compared. An independent samples t-test was conducted for this comparison. The normality assumption and the homogeneity of variance assumptions required for this parametric test were provided. The results of the examination of the assumptions are given in the [Supplementary](#) file. In addition, this study examined the invariance of the factor structure of the TSL-SES for Learners across genders. For this purpose, a partial invariance analysis was conducted.

### 2.3.5. Reliability estimations

The reliability of the scores obtained from TSL-SES for Learners were estimated via Cronbach  $\alpha$ , Guttman lambda 2 ( $\lambda$ ), latent class reliability coefficient (LCRC), composite reliability coefficients (CR), and McDonald's omega ( $\omega$ ).

## 2.4. Preparation of Data for Analyses

In preparing the data for analyses, individuals with aberrant items scores were identified primarily. Individuals that have aberrant item scores affect validity results negatively, especially in scale development and scale adaptation studies (Meijer & Nering, 1997; Rupp, 2013; Sijsma & Molenaar, 2002; Şengül Avşar, 2023). Removing individuals with aberrant item scores before the analyses increases the accuracy of the analyses for validity (Liu *et al.*, 2019; Şengül Avşar, 2023). Individuals with aberrant item scores are identified through person-fit statistics (PFS).

Parametric or nonparametric PFS are used to determine aberrant item scores. Individuals with aberrant item scores were determined according to nonparametric normed number of Guttman errors ( $G_N^P$ ) and van der Flier's  $U3^P$  PFS in this study. To calculate the nonparametric PFS within the scope of nonparametric IRT, the data must fit with the MHM. The necessary checks were made in this study and were shown in the [Supplementary](#) file. Based on these statistics, 57 individuals identified as having aberrant item scores were removed from the data set. The path followed while determining the data set to be analysed is shown in [Figure 1](#).



**Figure 1.** Data set preparation stage.

### 3. FINDINGS

#### 3.1. Validity Analyses

To determine the construct validity of TSL-SES for Learners, MHM and AISP analyses were conducted as nonparametric IRT methods. Additionally, EFA and assessments of convergent and discriminant validity were performed within the framework of CTT. Analyses were conducted based on the scores obtained from the draft scale consisting of 24 items.

##### 3.1.1. Findings of MHM and AISP

The first step in determining the psychometric properties of the scores obtained from the TSL-SES for Learners was to define the measured construct using exploratory methods. To do this, the data were analysed according to MHM.  $H$  estimated according to MHM and stated as the scalability coefficient, and the standard error ( $SE$ ) values of these values are shown in Table 2.

**Table 2.**  $H$  coefficients and  $SH$  values for TSL-SES for Learners items.

Items	$H$	$SE$	Items	$H$	$SE$
Item 1	0.57	0.03	Item 13	0.57	0.03
Item 2	0.54	0.03	Item 14	0.60	0.03
Item 3	0.52	0.03	Item 15	0.60	0.03
Item 4	0.54	0.03	Item 16	0.60	0.03
Item 5	0.54	0.03	Item 17	0.49	0.03
Item 6	0.58	0.03	Item 18	0.53	0.03
Item 7	0.60	0.03	Item 19	0.49	0.03
Item 8	0.58	0.03	Item 20	0.56	0.03
Item 9	0.58	0.03	Item 21	0.56	0.03
Item 10	0.58	0.03	Item 22	0.55	0.03
Item 11	0.58	0.03	Item 23	0.54	0.03
Item 12	0.48	0.03	Item 24	0.57	0.03
All scale	0.57	0.02			

$H$ : Scalability coefficients,  $SE$ : Standard error of  $H$

It was observed that the TSL-SES items and the total score were scaled in accordance with the MHM. After scaling according to MHM, AISP was conducted. AISP results are given in Table 3.

**Table 3.** AISP results of TSL-SES for Learners.

$c$	Number of Scales	Items in the Scales
0.30	1	All items
0.35	1	All items
0.40	1	All items
0.45	1	All items
0.50	2	12, 17, 19; other items
0.55*	3	1, 2, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 20, 21, 22, 23, 24; 3, 4, 5; 17, 18, 19
0.60*	4	1, 2, 15, 16, 21, 22, 23, 24; 7, 8, 9, 10, 11, 13, 14; 4, 5, 6; 17, 18

\*: Some items could not be scaled according to the Mokken model.

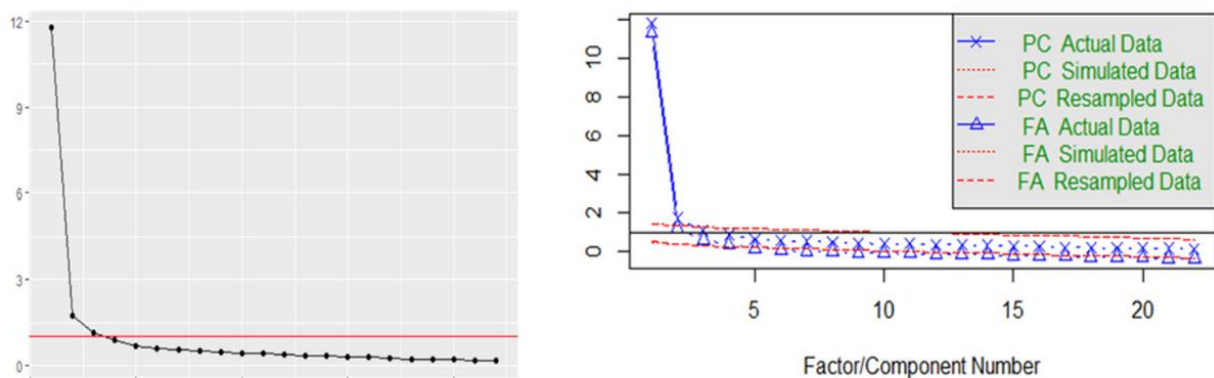
It was seen that a one-dimensional structure was generally indicated, however, a three-dimensional structure could be reached for 0.55, which was determined as a very high criterion for AISP estimations. AISP results indicated that there was a dominant factor in the scale according to  $c$  values 0.30 to 0.45.

### 3.1.2. Findings of EFA

Having determined that the item scores did not deviate significantly from the normal distribution, the factor structure was established using polychoric correlation with PAF. Since the number of items exceeded 10 times the sample size, and the KMO value and Bartlett's sphericity test results ( $KMO=0.96$ ,  $\chi^2=7838.41$ ,  $p<.05$ ) confirmed the sample size's adequacy for EFA. After determining that the necessary assumptions for EFA were provided, PAF was carried out.

In the first PAF result, Item 20 (*I can use TSL and spoken language together*) gave cross loadings of 0.36 on Factor 1, 0.38 on Factor 2; Item 24 (*I can achieve most of the goals I set for TSL*) gave overlapping loadings of 0.29 on Factor 1, 0.28 on Factor 2, and 0.37 on Factor 3, respectively. Thereupon, Item 24 was removed because it gave overlapping loadings on all three factors. The adequacy of the sample size was investigated again, and it was determined that the data set was suitable for EFA ( $KMO=0.96$ ,  $\chi^2=745.10$ ,  $p<.05$ ). In the results obtained, Item 20 gave an overlapping load of 0.37 on Factor 1 and 0.39 on Factor 2, again. It was decided to remove Item 20 as well.

The adequacy of the sample size was investigated again, and it was determined that the data was suitable for EFA ( $KMO=0.96$ ,  $\chi^2=7136.19$ ,  $p<.05$ ). Item-total correlation coefficients were calculated and shown in the [Supplementary](#) file. The lowest correlation coefficient between items was calculated as 0.20, while the highest correlation coefficient was calculated as 0.80. When deciding on the number of factors, three-factor proposals with eigenvalues greater than 1 (Kaiser, 1960), scree plot and parallel analysis results given in [Figure 2](#) were taken into consideration.



**Figure 2.** Scree plot and parallel analysis results.

When [Figure 2](#) is examined, both the PAF and parallel analysis results indicated a dominant single-factor structure. However, it should also be noted that a maximum of three-factor structure could be obtained from the parallel analysis results. In addition, it was determined that the eigenvalues obtained for three factors were 12.91, 1.73 and 1.09, respectively. The eigenvalues obtained for other factors were less than 1.

As a result, when PAF, parallel analysis and eigenvalues, and AISP results were evaluated together, a three-factor structure with a dominant factor was reached. [Table 4](#) shows item factor loadings and explained variance ratios for the three-factor structure.

[Table 4](#) shows that the lowest factor loadings for the three-factor structure are 0.46, which can be considered high. A three-factor structure with factor loadings at acceptable values was reached, and these three factors explained 67.00% of the total variance. When the items were examined in their factors, the first factor was called TSL Teaching Self-Efficacy, the second factor was called Communicate Effectively with TSL Self-Efficacy, and the third factor was called TSL Fluency/Application Self-Efficacy. The scale is given in the [Supplementary](#) file.



**Table 4.** Factor loadings\* and explained variance rates of TSL-SES for learners.

	Factor 1	Factor 2	Factor 3
Item1		0.83	
Item2		0.79	0.24
Item3	0.31	0.49	
Item4		0.74	
Item5	0.23	0.64	
Item6	0.32	0.49	
Item7	0.46	0.28	
Item8	0.79		
Item9	0.80		
Item10	0.88		
Item11	0.92		
Item12	0.73	0.23	0.39
Item13	0.68		
Item14	0.77		
Item15	0.55		0.34
Item16	0.61		0.41
Item17	0.70		
Item18	0.72		
Item19	0.61		
Item21		0.22	0.71
Item22	0.20		0.67
Item23		0.25	0.61
Variance explained (%)	0.37	0.19	0.11
Total variance	67.00		

\*: Factor loadings below 0.20 are not shown.

### 3.1.3 Findings of convergent and discriminant validity

The three-factor structure of the scale was tested with CFA in this study for convergent and discriminant validity proofs. Diagonally weighted least squares (DWLS) was used as the estimation method in CFA. The same data set was used on which EFA was conducted for CFA. This should be considered as a limitation. The obtained goodness of fit indices were as follows:  $\chi^2=1255.76$ ,  $df=209$ ,  $\chi^2/df=6.01$ , RMSEA=0.11, SRMR=0.07, CFI=0.98, TLI=0.98. Since the results were CFI>0.95, TLI>0.95 and SRMR<0.08, it could be stated that the scale structure was confirmed (Hair *et al.*, 2019).

AVE values were calculated as: 0.61 for Factor 1, 0.78 for Factor 2, for 0.82 Factor 3, respectively. These values exceed the 0.50 cutoff for AVE, indicating that convergent validity was achieved. The values obtained by calculating the square root of the AVE values were 0.78 for Factor 1, 0.88 for Factor 2, and 0.89 for Factor 3, respectively. Since these values were larger than the correlation values calculated between the factors ( $r_{f_1-f_2} = 0.78$ ,  $r_{f_1-f_3}=0.39$ ,  $r_{f_2-f_3}=0.48$ ), discriminant validity was ensured. HTMT values were calculated for discriminant validity. The obtained HTMT values (0.84, 0.68, and 0.73) were below the cutoff of 0.85, confirming that discriminant validity was achieved.

### 3.1.4. Findings of group differences

The scores obtained by those who had received sign language training before (n=200) were compared with the scores obtained by those who had not received sign language training (n=230). Accordingly, except for Factor 1 ( $t_{428}=1.22$ ,  $p>.05$ ), Factor 2 ( $t_{428}=3.02$ ,  $p<.05$ ); Factor 3 ( $t_{428}=3.85$ ,  $p<.05$ ), and the total score of the scale ( $t_{230}=31.29$ ,  $p<.05$ ) were found to be

statistically significant (the fact that the factor loadings are greater than 0.45, the evidence of convergent and discriminant validity, and high estimated reliability values indicate that total score can be obtained from the scale). These significant differences can be shown as complement/supportive evidence of construct validity.

The score differences between the first 27% of groups with the highest score (top group) and the last 27% of groups with the lowest score (bottom group) were also presented as supportive evidence for construct validity. Accordingly, Factor 1 ( $t_{230}=31.50, p<.05$ ); Factor 2 ( $t_{230}=31.61, p<.05$ ), Factor 3 ( $t_{230}=36.72, p<.05$ ), and the total score of the scale ( $t_{230}=31.29, p<.05$ ) were found to be statistically significant. These significant differences can be shown as complement/supportive evidence of construct validity.

This study also investigated measurement invariance across groups, but it was not achieved. In such cases where full invariance is not achieved, it is common to investigate partial invariance (Bryne, 2016; p. 231). In this study, the purpose of partial invariance was to demonstrate that the factor structure of the TSL-SES for Learners was similar across male and female groups. The fit indices obtained with the DWLS estimator were sufficient (CFI = 0.99, TLI = 0.99, RMSEA = 0.09, SRMR = 0.07), indicating that the factors were comparable across groups.

### 3.2. Reliability Analyses

The reliability coefficients obtained are presented in Table 5. In the event of an item being deleted, the item reliability statistics (for Cronbach's  $\alpha$  and McDonald's  $\omega$ ) to be obtained are provided for all factors in the Supplementary file.

**Table 5.** Reliability of scores obtained from TSL-SES for learners.

	Cronbach $\alpha$	Guttman 2 $\lambda$	LCRC	CR*	$\omega$
TSL Teaching Self-Efficacy	0.94	0.94	0.94	0.96	0.94
Communicate Effectively with TSL Self-Efficacy	0.89	0.89	0.89	0.90	0.89
TSL Fluency/Application Self-Efficacy	0.90	0.90	0.66	0.93	0.91
Total Scale	0.96	0.96	0.96	0.98	0.96

\*: It was calculated separately for the dimensions formed after EFA.

It is seen that the scores obtained from the TSL-SES for Learners yield reliable results. The values obtained from different reliability coefficients estimated in terms of internal consistency are very close to each other.

## 4. DISCUSSION and CONCLUSION

This study aimed to develop a scale giving valid and reliable results for measuring TSL-SES for Learners. The scale assesses self-efficacy in teaching self-efficacy, fluency/application self-efficacy, and effective communication self-efficacy for adult learners communicating with DHH individuals.

It was shown that TSL-SES for Learners gave valid and reliable results with various techniques in this study. Construct validity evidence was presented with different test theories. Evidence for the construct validity of the TSL-SES was presented with the IRT-based model-MHM together with the CTT-based model-EFA in this research. There are various studies that follow a similar method, especially in scale development studies (Chou *et al.*, 2017; Maldonado-Murciano *et al.*, 2024; Stochl *et al.*, 2012). It should also be noted that some studies (Blackwell *et al.*, 2023; Dagsdóttir *et al.*, 2023) developed the scales using only MSA. There are also other studies in the literature (Ommundsen *et al.*, 2007; Özberk *et al.*, 2021; Tran *et al.*, 2012) in which the factor structures of the scales were investigated according to Mokken analysis. AISP results showed that a three-factor structure could be achieved with a dominant factor in this study. Similar results were also obtained in EFA. This finding is similar to other research

findings where AISP and PAF gave similar results (de Cock *et al.*, 2011; Shenkin *et al.*, 2014; Şengül Avcı, 2022). This research is important in that it uses both CTT and NIRT models together in a scale development study and provides evidence of construct validity with more than one test theory.

In this study, one of the HTMT values estimated for discriminant validity was estimated as 0.84, just below the cut-off point of 0.85. In fact, this value is below the accepted limit in the literature, and it is even stated that the cut-off value can be expanded to 0.90 (Franke & Sarstedt, 2019; Henseler *et al.*, 2015). However, when the general acceptance is taken into consideration, it can be said that although it is technically accepted, it is relatively high. The reason for this can be thought to be that the scale has a dominant factor.

The TSL-SES for Learners was a 22-item scale with a five-category score, and there were no reverse-coded items. The lowest possible score was 22, and the highest was 110. The first dimension of the scale, "TSL Teaching Self-Efficacy," consisted of 13 items; the second dimension, "Communicate Effectively with TSL Self-Efficacy," consisted of six items; and the third dimension, "TSL Fluency/Application Self-Efficacy," consisted of three items.

In addition, it was determined that the scores obtained from TSL-SES for Learners gave reliable results when their estimates were evaluated with different reliability coefficients. TSL-SES for Learners was a scale with high internal consistency.

Self-Efficacy in Teaching-Factor 1 of TSL focuses on the skills of preparation, material preparation, classroom management, mastering the features of the language, structuring the lesson, developing effective assessment tools and preparing accessible learning content. There are research findings related to sign language proficiency in the literature supporting this dimension. Bintoro *et al.* (2023), who stated that teachers should evaluate the learning needs of DHH students in the most accurate way when teaching sign language, drew attention to the fact that they used three basic strategies in teaching. They reported that teachers increased the participation of DHH students in the classroom by using sign language, verbal communication and total communication methods consisting of both sign language and verbal spoken language. There are also research examples that emphasize the features such as material, teaching structure and classroom management that stand out in Factor 1. For example, De Picker (2020) emphasized that teachers had difficulties in conducting classroom activities in classes attended by DHH students. Kelly *et al.* (2020) pointed out that DHH students had difficulties in communicating with their teachers. Students reported that their teachers could not understand their educational needs in terms of communication self-efficacy. Therefore, it was suggested that teachers' competence in understanding and teaching DHH students should be supported.

The studies under consideration emphasise the necessity of inclusive educational environments for children with hearing impairments (DHH), highlighting that this is a legal right. They further recommend the widespread implementation of sign language in educational institutions, the mandatory nature of sign language courses for all students, and the clarification of the roles and responsibilities of interpreters.

In the project Promoting Excellence in Sign Language Instruction (2016-2019) of the European Centre for Modern Languages (ECML, 2024), examinations were made on what kind of competencies sign language speakers should master. In these examinations, it was pointed out that there should be standards in sign language education, and guidelines should be prepared on which competencies and characteristics sign language instructors should have.

The present study aims to determine the self-efficacy related to the dimension of communication with TSL in terms of using sign language actively in the communication process, communicating with DHH individuals, using sign language in TSL lessons, making gestures and finger shapes simultaneously, and using different teaching methods and technology tools effectively. This is achieved by means of the Self-Efficacy to Communicate Effectively with TSL-Factor 2. The self-efficacy and effective communication skills that come

to the fore in this factor are like the findings of the studies in the literature examining sign language self-efficacy. Reddy (2011) adapted the English general self-efficacy scale into ASL for DHH individuals, maintaining its original psychometric structure. While confirming scale validity, the study highlighted the need to examine predictors of self-efficacy in DHH populations.

TSL Fluency and Application Self-Efficacy-Factor 3 consisted of self-reports that aim to determine the competence to use TSL fluently, to understand the TSL interpreter, and to explain a Turkish text in TSL. A similar study was conducted in the development of ASL Self-Efficacy (Koch, 2023). This study emphasized the importance of feedback and its use in teaching for the development of ASL self-efficacy of 7th-grade students. The comprehension of the TSL communicator by the DHH individual or the TSL learner will be an effective determinant of the person's TSL fluency self-efficacy. Studies have explored the efficacy of teachers working with DHH students. Garberoglio *et al.* (2012) examined teachers' sense of efficacy and its impact on DHH students' performance, finding significant differences in efficacy beliefs related to student engagement, instructional strategies, and classroom management. Graham *et al.* (2021) focused on primary school teachers of DHH students, using Bandura's theory to investigate their self-efficacy in teaching writing. The study highlighted teachers' beliefs about writing, including self-efficacy, attitudes, and epistemological beliefs, and found that teachers demonstrated strong efficacy in supporting the development of DHH students' writing skills.

Providing valid and reliable results for TSL learners, TSL-SES allows the self-efficacy levels of individuals to be examined throughout their learning process. This can enable the development of effective interventions for teaching TSL to individuals or groups with low self-efficacy despite knowing TSL. By analyzing the communication patterns of individuals with high TSL self-efficacy, existing TSL communication strategies can inform teaching plans. Given that perceived self-efficacy strongly influences behavior, effort, resilience, and goal-setting, assessing the TSL self-efficacy of personnel working with DHH individuals is important.

This study has some limitations. Although evidence is provided for the construct validity of the scale developed in the study with different test theories, it would be appropriate to monitor the scale and estimate the construct validity proofs with CFA with different groups. Additionally, although the factor structure of the scale was determined to be similar across gender groups with partial invariance, measurement invariance can be examined with multigroup CFA with different samples. In addition, test-retest reliability can be used in reliability estimates.

The combined use of AISP and PAF in investigating the construct validity of TSL-SES for Learners yielded effective results in revealing the factor structure. It may be recommended to follow this method in scale development studies. TSL-SES for Learners developed is for Turkish-Sign-Language. However, the scale can be adapted for different sign languages for self-efficacy.

### Acknowledgments

The authors thank to valuable contribution to all participants to this research.

**Data availability:** The data used in this research can be shared if upon reasonable request and with permission of all authors of this research.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. The study was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2001). **Ethics Committee Number:** Recep Tayyip Erdoğan University, the Ethics Committee, (Document No. 147, June 14, 2022).

## Contribution of Authors

**P.P.A., N.K., N.I.U., A.Z.O.:** Conceptualization. **A.Ş.A.:** Methodology, Formal analysis and data visualisation. **P.P.A., N.K., N.I.U., A.Z.O., A.Ş.A.:** Writing-original draft preparation. **A.Ş.A., N.K., N.I.U.:** Writing-review and editing. **N.K., N.I.U.:** Supervision. All authors have read and agreed to the uploaded manuscript.


## Orcid

Pelin Piştav Akmeşe  <https://orcid.org/0000-0001-8269-3899>

Asiye Şengül Avşar  <https://orcid.org/0000-0001-5522-2514>

Nilay Kayhan  <https://orcid.org/0000-0002-0937-8013>

Necla Işıkdوغان Uğurlu  <https://orcid.org/0000-0002-1795-0470>

Ayşen Zeynep Oral  <https://orcid.org/0000-0001-6378-5464>

## REFERENCES

- Alfano, A.R., Radlinski, S., & del Corro-Helbig, M.G. (2022). Challenges and opportunities with deaf multilingual learners. (Eds. Musyoka, M.M.) *Deaf education and challenges for bilingual/multilingual students*, 1-39. IGI Global Scientific Publishing.
- Almayez, M.A., Al-Khresheh, M.H., Al-Qadri, A.H., Alkhateeb, I.A., & Alomaim, T.I.M. (2025). Motivation and English self-efficacy in online learning applications among Saudi EFL learners: Exploring the mediating role of self-regulated learning strategies. *Acta Psychologica*, 254, 104796. <https://doi.org/10.1016/j.actpsy.2025.104796>
- Anastasia, A. (1976). *Psychological testing* (6th ed.). Macmillan Publishing.
- Artino, A.R. (2012). Academic self-efficacy: from educational theory to instructional practice. *Perspect Med Educ*, 1, 76–85. <https://doi.org/10.1007/s40037-012-0012-5>
- Avşar, V., & Sevim, S.A. (2022). The effectiveness of cognitive behavioral therapy including updating the early life experiences and images with the empty chair technique on social anxiety. *International Journal of Assessment Tools in Education*, 9(1), 181-202. <https://doi.org/10.21449/ijate.1062613>
- Ballenger, S. (2025). Effective communication for deaf and hard of hearing people. Center for inclusive design and innovation. Georgia Tech-College of Design. [https://accessga.gatech.edu/sites/default/files/2025-04/Effective%20Communication%20for%20Deaf%20and%20Hard%20of%20Hearing%20People\\_Final.pdf](https://accessga.gatech.edu/sites/default/files/2025-04/Effective%20Communication%20for%20Deaf%20and%20Hard%20of%20Hearing%20People_Final.pdf)
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman.
- Baumeister, R.F., & Vohs, K.D. (2007). Self-efficacy. In *Encyclopedia of Social Psychology* (pp. 815-817). SAGE Publications, Inc. <https://doi.org/10.4135/9781412956253>
- Beal, J.S., Dostal, H.M., & Easterbrooks, S.R. (2024). *Literacy instruction for students who are deaf and hard of hearing: (2nd ed.) (Perspectives on Deafness)*. Oxford University Press.
- Bilgiç, H.C., Aslan, C., Özdemir Kılıç, C., & Kan, A. (2021). Developing an attitude scale for Turkish-Sign-Language (TSL-AS). *Participatory Educational Research*, 8(1), 200-218. <http://dx.doi.org/10.17275/per.21.11.8.1>
- Bintoro, T., Fahrurrozi, Kusmawati, A.P., & Dewi, R.S. (2023). The teacher strategies in teaching sign language for deaf students in special schools Jakarta. *Cogent Education*, 10(2). <https://doi.org/10.1080/2331186X.2023.2258294>
- Blackwell, C.K., Sherlock, P., Jackson, K.L., Hofheimer, J.A., Cella, D., Algermissen, M.A., Alshawabkeh, A.N., Avalos, L.A., Bastain, T., Blair, C., Bosquet Enlow, M., Brennan, P. A., Breton, C., Bush, N.R., Chandran, A., Collazo, S., Conradt, E., Crowell, S.E., Deoni, S., . . . & Margolis, A.E. (2023). Development and psychometric validation of the Pandemic-Related Traumatic Stress Scale for children and adults. *Psychological Assessment*, 35(11), 1054-1067. <https://doi.org/10.1037/pas0001211>
- Bryne, B. (2016). *Structural equation modeling with amos basic concepts, applications, and programming*. (3rd ed.). Routledge.



- Chen, Y., Watson, R., & Hilton, A. (2016). An exploration of the structure of mentors' behavior in nursing education using exploratory factor analysis and Mokken scale analysis. *Nurse Education Today*, 40, 161-167. <https://doi.org/10.1016/j.nedt.2016.03.001>
- Chou, Y.H., Lee, C.P., Liu, C.Y., & Hung, C.I. (2017). Construct validity of the Depression and Somatic Symptoms Scale: evaluation by Mokken scale analysis. *Neuropsychiatric Disease and Treatment*, 13, 205-211. <https://doi.org/10.2147/NDT.S118825>
- Cohen, R.J., Swerdlik, M., & Sturman, E.D. (2013). *Psychological testing and assessment: An introduction to tests and measurement* (8th ed.). McGraw-Hill Companies.
- Council of Higher Education (2018). Özel eğitim öğretmenliği lisans programı [Special education teaching undergraduate program]. [https://eski.yok.gov.tr/Documents/Kurumsal/egitim\\_ogretim\\_dairesi/Yeni-Ogretmen-Yetistirme-Lisans-Programlari/Ozel\\_Egitim\\_Ogretmenligi\\_Lisans\\_Programi.pdf](https://eski.yok.gov.tr/Documents/Kurumsal/egitim_ogretim_dairesi/Yeni-Ogretmen-Yetistirme-Lisans-Programlari/Ozel_Egitim_Ogretmenligi_Lisans_Programi.pdf)
- Crişan, D.R., Tendeiro, J.N., & Meijer, R.R. (2022). The Crit coefficient in Mokken scale analysis: A simulation study and an application in quality-of-life research. *Quality of Life Research*, 31(1), 49-59. <https://doi.org/10.1007/s11136-021-02924-z>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Cengage Learning.
- Dagsdóttir, B.E., Kristjánisdóttir, H., Vésteinsdóttir, V., & Thorsdóttir, F. (2023). Task and ego orientation in sport questionnaire: A Mokken scale analysis. *Sage Open*, 13(3). <https://doi.org/10.1177/21582440231195200>
- Danermark, B., Antonson, S., & Lundström, I. (2001). Social inclusion and career development-transition from upper secondary school to work or post-secondary education among hard of hearing students. *Scandinavian Audiology*, 30(2), 120-128. <https://doi.org/10.1080/010503901750166880>
- Darıca, N., & Şipal, F. (2011). İşitme engelli çocuklarda gelişim ve eğitsel müdahale [Development and educational intervention in hearing impaired children]. Hacettepe University Press.
- de Cock, E.S., Emons, W.H., Nefs, G., Pop, V.J., & Pouwer, F. (2011). Dimensionality and scale properties of the Edinburgh Depression Scale (EDS) in patients with type 2 diabetes mellitus: the DiaDDzoB study. *BMC Psychiatry*, 11(1), 141. <https://doi.org/10.1186/1471-244X-11-141>
- De Picker, M. (2020). Rethinking inclusion and disability activism at academic conferences: Strategies proposed by a PhD student with a physical disability. *Disability & Society*, 35(1), 163-167. <https://doi.org/10.1080/09687599.2019.1619234>
- DeVellis, R.F. (2017). *Scale development: Theory and applications* (4th ed.). Sage publications.
- Dıkyuva, H., Makaroğlu, B., & Arık, E. (2015). Türk İşaret Dili dilbilgisi kitabı [Turkish-Sign-Language grammar book]. Ministry of Family and Social Policies Publications.
- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, 23(3), 225-241. <https://doi.org/10.1177/073428290502300303>
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.
- Emons, W.H., Sijtsma, K., & Pedersen, S.S. (2012). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients: comparison of Mokken scale analysis and factor analysis. *Assessment*, 19(3), <https://doi.org/10.1177/1073191110384951>
- Erickson, G., & Noonan, P.M. (2025). Selfefficacy assessment suite: Technical report. College & Career Competency Framework. <https://www.cccframework.org/wp-content/uploads/Self-EfficacyAssessSuiteTech.pdf>
- Eryiğit, C. (2017). *Yazılı Türkçe dilinden Türk İşaret Diline (TİD) makine çevirisi sistemi* [Text to sign language machine translation system for Turkish] [Unpublished doctoral dissertation]. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

- European Centre for Modern Languages (ECML). (2024, May 18). Promoting excellence in sign language instruction Instructions for learners-Assessment. Retrieved June 25, 2025, from [www.ecml.at/prosign](http://www.ecml.at/prosign)
- Fitriyani, F., Ainii, L.Q., Jannah, R., & Maryam, S. (2024). Analysis of sign language skills in improving communication and learning for deaf children. *Continuous Education: Journal of Science and Research*, 5(1), 30-39. <https://doi.org/10.51178/ce.v5i1.1757>
- Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50. <https://doi.org/10.2307/3151312>
- Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education*. (8th ed.). McGraw-Hill.
- Franke, G., & Sarstedt, M. (2019). Heuristics versus statistics in discriminant validity testing: a comparison of four procedures. *Internet Research*, 29(3), 430-447. <https://doi.org/10.1108/IntR-12-2017-0515>
- Galindo-Garre, F., Hendriks, S.A., Volicer, L., Smalbrugge, M., Hertogh, C.M., & van der Steen, J.T. (2014). The Bedford Alzheimer nursing-severity scale to assess dementia severity in advanced dementia: a nonparametric item response analysis and a study of its psychometric characteristics. *American Journal of Alzheimer's Disease & Other Dementias®*, 29(1), 84-89. <https://doi.org/10.1177/1533317513506777>
- Gao, S., Ma, X., Tsui, H., Wang, J., & Zhang, X. (2024). Item response theory analysis of the Chinese version compulsive shopping scale. *Comprehensive Psychiatry*, 135, 152535. <https://doi.org/10.1016/j.comppsy.2024.152535>
- Garberoglio, C.L., Gobble, M.E., & Cawthon, S.W. (2012). A national perspective on teachers' efficacy beliefs in deaf education. *The Journal of Deaf Studies and Deaf Education*, 17(3), 367-383. <https://doi.org/10.1093/deafed/ens014>
- Graham, S., Wolbers, K., Dostal, H., & Holcomb, L. (2021). Does teacher self-efficacy predict writing practices of teachers of deaf and hard of hearing students? *Journal of Deaf Studies and Deaf Education*, 26(3), 438-450. <https://doi.org/10.1093/deafed/enab012>
- Guardino, C. (2015). Evaluating teachers' preparedness to work with students who are deaf and hard of hearing with disabilities. *American Annals of the Deaf*, 160(4), 415-426. <https://www.jstor.org/stable/26235231>
- Gürboğa, C., & Kargın T. (2003). İşitme engelli yetişkinlerin farklı ortamlarda kullandıkları iletişim yöntemlerinin/becerilerinin incelenmesi [Examining the communication methods/skills used by hearing impaired adults in different environments]. *Ankara University Journal of Faculty of Educational Sciences (JFES)*, 36(1), 51-64. [https://doi.org/10.1501/Egifak\\_0000000074](https://doi.org/10.1501/Egifak_0000000074)
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2019). *Multivariate data analysis* (8th ed.). Annabel Ainscow.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage Publications.
- Hammer, S.G. (1998). The cost of treating deaf and hard-of-hearing patients. *American Family Physician*, 58(3), 659.
- Henseler, J., Ringle, C.M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43, 115-135. <https://doi.org/10.1007/s11747-014-0403-8>
- Humphries, T., Mathur, G., Napoli, D.J., Padden, C., & Rathmann, C. (2022). Deaf children need rich language input from the start: Support in advising parents. *Children*, 9(11):1609. <https://doi.org/10.3390/children9111609>
- Jones, B.E., Clark, G.M., & Soltz, D.F. (1997). Characteristics and practices of sign language interpreters in inclusive education programs. *Exceptional Children*, 63(2), 257-268. <https://doi.org/10.1177/001440299706300209>

- Jorgensen, T.D., Pornprasertmanit, S., Schoemann, A.M., Rosseel, Y., Miller, P., Quick, C., Garnier-Villareal, M., Selig, J., Boulton, A., Preacher, K., Coffman, D., Rhemtulla, M., Robitzsch, A., Enders, C., Arslan, R., Bell, C., Panko, P., Merkle, E., Chesnut, S., ... Vanbrabant, L. (2023). *semTools: Useful tools for structural equation modeling* (Version 0.5-7) [R package]. Comprehensive R Archive Network. <https://doi.org/10.32614/CRAN.package.semTools>
- Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1), 141-151. <https://doi.org/10.1177/00131644600200116>
- Kelly, J.F., McKinney, E.L., & Swift, O. (2020). Strengthening teacher education to support deaf learners. *International Journal of Inclusive Education*, 26(13), 1289–1307. <https://doi.org/10.1080/13603116.2020.1806366>
- Kim, H. (2024). Exploring the interplay of language mindsets, self-efficacy, engagement, and perceived proficiency in L2 learning. *Humanit Soc Sci Commun* 11, 1295. <https://doi.org/10.1057/s41599-024-03783-y>
- Kim, H.Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, 38(1), 52-54. <https://doi.org/10.5395/rde.2013.38.1.52>
- Kim, S.J., Kwon, M.S., & Han, W. (2015). Development of a school adaptation program for elementary school students with hearing impairment. *Journal of Audiology & Otology*, 19(1), 26. <http://dx.doi.org/10.7874/jao.2015.19.1.26>
- Kline, R.B. (2011). *Principles and practice of structural equation modeling*. Guilford Publications.
- Koch, K.E. (2023). Increasing students' self-efficacy in American sign language using timely and specific instructional feedback. (Unpublished master's thesis). Wilmington University.
- Koopman, L., Zijlstra, B.J., & Van der Ark, L.A. (2023). Evaluating model fit in two-level mokken scale analysis. *Psych*, 5(3), 847-865. <https://doi.org/10.3390/psych5030056>
- Kubuş, O., İlkbaşaran, D., & Gilchrist, S. (2016). Türkiye'de işaret dili planlaması ve Türk İşaret Dili'nin yasal durumu [Sign language planning in Türkiye and the legal situation of Turkish-Sign-Language]. In E. Arık (Ed.), *Ellerle konuşmak: Türk İşaret Dili araştırmaları* [Talking with Hands: Turkish-Sign-Language Research] (pp. 23–50). Koç University.
- Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The impact of aberrant response on reliability and validity. *Measurement: Interdisciplinary Research and Perspectives*, 17(3), 133-142. <https://doi.org/10.1080/15366367.2019.1584848>
- Ma, Y. (2022). The effect of teachers' self-efficacy and creativity on English as a foreign language learners' academic achievement. *Frontiers in Psychology*, 13, 872147. <https://doi.org/10.3389/fpsyg.2022.872147>
- Maldonado-Murciano, L., Pontes, H.M., Barrios, M., Gómez-Benito, J., & Guilera, G. (2024). Mokken scale analysis of the Internet Gaming Disorder Scale–Short-Form and the Gaming Disorder Test. *Addictive Behaviors Reports*, 100567. <https://doi.org/10.1016/j.abrep.2024.100567>
- Marriage, J., Brown, T.H., & Austin, N. (2017). Hearing impairment in children. *Paediatrics and Child Health*, 27(10), 441-446. <https://doi.org/10.1016/j.paed.2017.06.003>
- McDermid, C. (2020). Educational interpreters, deaf students and inclusive education? *Turkish Journal of Special Education Research and Practice*, 2(1), 27-46. <https://doi.org/10.37233/TRSPED.2020.0107>
- Meijer, R.R., & Nering, M.L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement*, 21(4), 321-336. <https://doi.org/10.1177/01466216970214003>
- Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: an expanded source book*. SAGE Publications.

- MoNE (Milli Eğitim Bakanlığı) (2018). Özel Eğitim Hizmetleri Yönetmeliği [Special Education Services Regulation]. [https://orgm.meb.gov.tr/meb\\_iys\\_dosyalar/2018\\_07/09101900\\_ozel\\_egitim\\_hizmetleri\\_yonetmeli\\_07072018.pdf](https://orgm.meb.gov.tr/meb_iys_dosyalar/2018_07/09101900_ozel_egitim_hizmetleri_yonetmeli_07072018.pdf)
- Murray, J.J., Hall, W.C., & Snoddon, K. (2019). Education and health of children with hearing loss: the necessity of signed languages. *Bulletin of the World Health Organization*, 97(10), 711.
- Musyoka, M.M. (2022). *Deaf education and challenges for bilingual/multilingual students*. IGI Global Scientific Publishing.
- NDC (National Deaf Center) (2019). Importance of effective communication between deaf and hearing individuals. Retrieved June 18, 2024, from <https://nationaldeafcenter.org/wp-content/uploads/2022/11/Importance-of-Effective-Communication-Between-Deaf-and-Hearing-Individuals.pdf>
- Nikolarazi, M., & Argyropoulos, V. (2015). The learning and communication barriers of deaf and hard of hearing students in higher education. In *EDULEARN15 Proceedings* (pp. 4130-4134). IATED.
- Ommundsen, R., van der Veer, K., Van Le, H., Krumov, K., & Larsen, K.S. (2007). Developing attitude statements toward illegal immigration: Transcultural reliability and utility. *Psychological Reports*, 100(3), 901-914. <https://doi.org/10.2466/pr0.100.3.901-914>
- Özberk, E.H., Ünsal Özberk, E.B., Uluç, S., & Öktem, F. (2021). Investigating invariant item ordering in intelligence tests: Mokken scale analysis of KBIT-2. *International Journal of Assessment Tools in Education*, 8(3), 714-728. <https://doi.org/10.21449/ijate.858183>
- Palm, K.M., & Strong, D.R. (2007). Using item response theory to examine the white bear suppression inventory. *Personality and Individual Differences*, 42(1), 87-98. <https://doi.org/10.1016/j.paid.2006.06.023>
- Panjie, D., & Velarde, J. (2025). A Study on the correlation between language self-efficacy and language learning strategies of non-english majors. *International Journal of Academic Research in Progressive Education and Development*, 14(2), 635-646. <http://dx.doi.org/10.6007/IJARPED/v14-i2/24961>
- Pfitzner-Eden, F. (2016). Why do I feel more confident? Bandura's sources predict preservice teachers' latent changes in teacher self-efficacy. *Frontiers in Psychology*, 7, 1486. <https://doi.org/10.3389/fpsyg.2016.01486>
- Piştav Akmeşe, P. (2019). Eğitimde Türk İşaret Dili [Turkish-Sign-Language in education]. Nobel Yayıncılık.
- Piştav Akmeşe, P., Kayhan, N., Kirazlı, G., Öğüt, F., & Kirazlı, T. (2018). Odyoloji ve konuşma bozuklukları alanında lisansüstü eğitim öğrencilerinin işitme kayıplı çocukların dil, konuşma ve iletişim becerilerinin desteklenmesi ve eğitimleri hakkındaki görüşleri. *Türkiye Sağlık Bilimleri ve Araştırmaları Dergisi*, 1(1), 13-23.
- Pretorius, T.B., & Padmanabhanunni, A.A. (2024). Unidimensional short form of the Beck Hopelessness Scale (BHS-7) derived using item response theory. *Scientific Reports*, 14, Article number: 6021. <https://doi.org/10.1038/s41598-024-56792-x>
- Reddy, M.Z. (2011). The development of an American sign language general self-efficacy scale for use with deaf individuals [Unpublished doctoral dissertation]. Alliant International University.
- Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version 2.3.3) [R package]. Comprehensive R Archive Network. <https://cran.r-project.org/web/packages/psych/index.html>
- Rosseel, Y., Jorgensen, T.D., De Wilde, L., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Rockwood, N., Scharf, F., Du, H., Jamil, H., & Classe, F. (2023). *lavaan: Latent variable analysis* (Version 0.6-15) [R package]. Comprehensive R Archive Network. <https://cran.r-project.org/package=lavaan>



- Rupp, A.A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, 55(1), 3.
- Salehomoum, M. (2020). Inclusion of signing deaf or hard-of-hearing students: factors that facilitate versus challenge access and participation. *Perspectives of the ASHA Special Interest Groups*, 5(4), 971-983. [https://doi.org/10.1044/2020\\_PERSP-19-00124](https://doi.org/10.1044/2020_PERSP-19-00124)
- Scott, J.A., & Hoffmeister, R.J. (2017). American Sign Language and academic English: Factors influencing the reading of bilingual secondary school deaf and hard of hearing students. *The Journal of Deaf Studies and Deaf Education*, 22(1), 59-71. <https://doi.org/10.1093/deafed/enw065>
- Shenkin, S.D., Watson, R., Laidlaw, K., Starr, J.M., & Deary, I.J. (2014). The attitudes to ageing questionnaire: Mokken scaling analysis. *PLOS ONE* 9(9), e108766. <https://doi.org/10.1371/journal.pone.0099100>
- Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Sage Publications.
- Stochl, J. (2007). Nonparametric extension of item response theory models and its usefulness for assessment of dimensionality of motor tests. *Acta Universitatis Carolinae*, 42(1), 75-94.
- Stochl, J., Jones, P.B., & Croudace, T.J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: A non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology*, 12(1), 1-16. <https://doi.org/10.1186/1471-2288-12-74>
- Sun, Z., & Mu, B. (2023). Motivating online language learning: exploring ideal L2 self, grit, and self-efficacy in relation to student satisfaction. *Frontiers in Psychology*, 14, 1293242. <https://doi.org/10.3389/fpsyg.2023.1293242>
- Szarkowski, A., Moeller, M.P., Gale, E., Smith, T., Birdsey, B.C., Moodie, S.T., ... & Zheng, X. (2024). Family-centered early intervention deaf/hard of hearing (FCEI-DHH): Cultural & Global Implications. *The Journal of Deaf Studies and Deaf Education*, 29(SI), SI27-SI39. <https://doi.org/10.1093/deafed/enad036>
- Şengül Avşar, A. (2022). Comparing the automatic item selection procedure and exploratory factor analysis in determining factor structure. *Participatory Educational Research*, 9(2), 416-436. <https://doi.org/10.17275/per.22.47.9.2>
- Şengül Avşar, A. (2023). Aberrant individuals' effects on fit indices both of confirmatory factor analysis and polytomous IRT models. *Current Psychology*, 42(3), 2157-2166. <https://doi.org/10.1007/s12144-021-01563-4>
- Şengül Avşar, A., & Barış Pekmezci, F. (2022). Examination of motivation scales: Is the purpose academic promotion or the need to measure psychological constructs? *Psycho-Educational Research Reviews*, 11(3), 774-791. [https://doi.org/10.52963/PERR\\_Biruni\\_V11.N3.19](https://doi.org/10.52963/PERR_Biruni_V11.N3.19)
- Tang, G. (2024). Sign language and inclusive deaf education: An Asian perspective. *Deafness & Education International*, 26(1), 1–5. <https://doi.org/10.1080/14643154.2024.2302702>
- Tendeiro, J.N. (2023). *PerFit: Person Fit* (Version 1.3.0) [R package]. Comprehensive R Archive Network. <https://cran.r-project.org/web/packages/PerFit/PerFit.pdf>
- Tran, U.S., Stieger, S., & Voracek, M. (2012). Psychometric analysis of Stöber's Social Desirability Scale (SDS-17): An item response theory perspective. *Psychological Reports*, 111(3), 870-884. <https://doi.org/10.2466/03.09.PR0.111.6.870-884>
- Tuohimaa, K., Loukusa, S., Löppönen, H., Välimaa, T., & Kunnari, S. (2022). Communication abilities in children with hearing loss—views of parents and daycare professionals. *Journal of Communication Disorders*, 99, 106256. <https://doi.org/10.1016/j.jcomdis.2022.106256>
- Ubido, J., Huntington, J., & Warburton, D. (2002). Inequalities in access to healthcare faced by women who are deaf. *Health & Social Care in the Community*, 10(4), 247-253. <https://doi.org/10.1046/j.1365-2524.2002.00365.x>



- Usher, E.L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, 78(4), 751-796. <https://doi.org/10.3102/0034654308321456>
- Van der Ark, L.A., Koopman L., Straat, H., & Van Den B. (2023). *Mokken: Mokken Scale Analysis in R* (Version 3.0.6). [R package]. <https://cran.r-project.org/web/packages/mokken/mokken.pdf>
- Wilson-Menzfeld, G., Gates, J.R., Jackson-Corbett, C., & Erfani, G. (2025). Communication experiences of deaf/hard-of-hearing patients during healthcare access and consultation: a systematic narrative review. *Health & Social Care in the Community*, 2025(1), 8867224. <https://doi.org/10.1155/hsc/8867224>

## The development of the Sustainable Consumption Behavior and Intention Scale

Merve Eker Çelebi<sup>1\*</sup>, Fatma Taşkın Ekici<sup>1</sup>

<sup>1</sup>Pamukkale University, Faculty of Education, Department of Mathematics and Science Education, Denizli, Türkiye

### ARTICLE HISTORY

Received: May 6, 2024

Accepted: July 26, 2025

### Keywords:

5R,  
Sustainable consumption  
behavior,  
Sustainable intention,  
Zero waste.

**Abstract:** The aim of this study is to develop a measurement tool entitled the Sustainable Consumption Behavior and Intention Scale (SuCBIS) that provides valid and reliable data for identifying individuals' sustainable consumption behaviors and their intentions within the framework of zero waste regulation. The scale was constructed based on the steps of the zero-waste hierarchy: reconsider, reuse, recycle, and rot (compost). The research was designed within the scope of a descriptive survey model based on scale development. For the validation and reliability processes of the scale, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were employed. Convenience sampling was used in the study. The EFA was conducted with 242 pre-service teachers, while the CFA was carried out with a different sample of 280 pre-service teachers. The participants were undergraduate students enrolled in various teacher education programs. Content validity was ensured through literature review, the opinions of pre-service teachers, and expert evaluations. As a result of EFA, a five-factor structure was revealed regarding consumption behavior: reconsider, reuse, reduce, recycle, and compost. CFA findings indicated that the developed model showed an acceptable level of model-data fit. The reliability of the data obtained from the scale was supported by internal consistency analyses and item-total correlations. In conclusion, a measurement tool that provides valid and reliable data for evaluating pre-service teachers' sustainable consumption behaviors and intentions has been developed.

## 1. INTRODUCTION

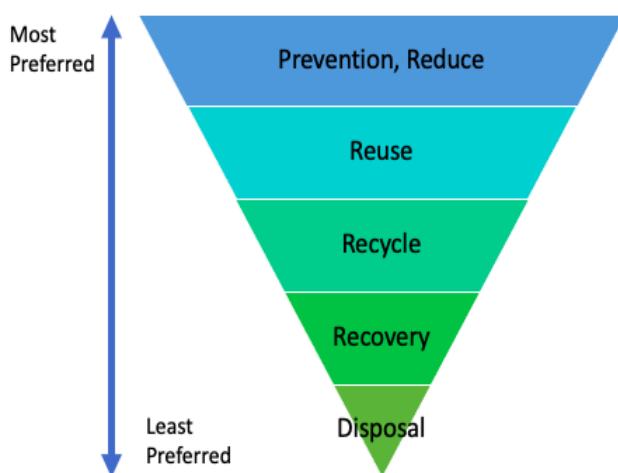
Reducing waste on Earth is among the important issues for the future and the sustainability of the world. Zero waste is defined as a waste management process and approach that includes preventing waste in the evaluation of waste, using resources more efficiently, preventing or minimizing waste generation, and separating and recycling waste at its source if it occurs (Zero Waste Regulation, 2019: Article 4). Zero waste is closely related to issues such as limited natural resources and pollution. It is an approach that aims to protect the environment and human health and all resources by preventing/reducing waste generation in production, consumption and service processes, prioritizing reuse, collecting the generated waste separately at the source, and reducing the amount of waste to be sent for disposal by ensuring recycling

\*CONTACT: Merve EKER ÇELEBİ ✉ [mervee@pau.edu.tr](mailto:mervee@pau.edu.tr) 📠 Department of Mathematics and Science Education, Denizli, Türkiye

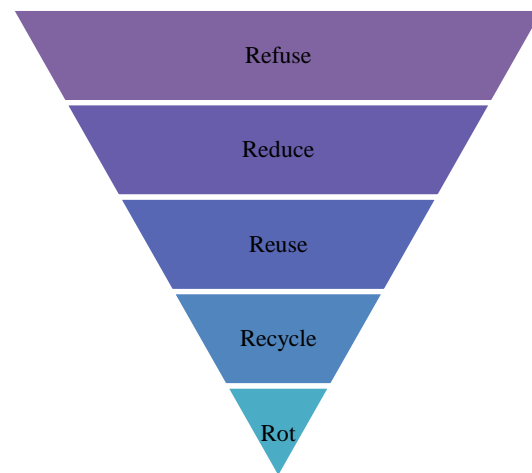
The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

and/or recovery (Bektaş, 2020). The reason for adopting the zero-waste approach in many countries is that it encourages sustainable production and consumption, sustainable development, recycling and resource transformation (Zaman, 2015).

Along with the zero-waste goal, environmentally friendly consumption habits also support sustainability and waste reduction efforts. Environmentally friendly consumption includes the understanding of choosing nature-friendly products and using resources efficiently in order to reduce environmental impacts. Environmentally friendly products can be products made from renewable and recyclable materials or those that consume low energy. In addition, extending the life of products, encouraging reuse and using recycling systems effectively are examples of environmentally friendly consumption. According to the zero-waste regulation, the most priority option according to the waste hierarchy is considered as prevention and reduction, as shown in Figure 1. The last option is disposal, and in this process, energy recovery is stated as the last method that should be preferred according to the waste management hierarchy (Bektaş, 2020).



**Figure 1.** Methods of dealing with waste according to waste hierarchy (Bektaş, 2020).



**Figure 2.** Zero waste process with 5R code of conduct (Johnson, 2019, p.13).

Johnson (2019) visualized the approach in the zero-waste process in Figure 2. In Figure 1 and Figure 2, reduction, reuse and recycling are seen among the commonalities in waste reduction methods. The zero-waste approach includes the 3R rule of "Reduction, Reuse and Recycle", which forms the basis of environmental awareness (Song *et al.*, 2015). When composting is associated with disposal, the only difference between the two images may be the extent of energy recovery.

In order to achieve the zero waste goal, "Refuse", derived from the English word starting with the letter R, means thinking carefully before purchasing and rejecting if it is not really needed; "Reduce", living more simply and reducing what you do not need; "Reuse" to use for different purposes or to use repeatedly; "Recycle" means, first of all, calculating that you will recycle when buying things that you cannot reject, reduce or reuse; The fifth and last dimension is considered as "Rot" (Johnson, 2019).

If people change their consumption behavior and adopt a lifestyle in accordance with the 5R behavior, waste production can be minimized or approach zero. Adopting this approach, Johnson (2019) explained it with examples from his life in his book *Zero Waste Home*. She explains many of the methods she has implemented in her life, such as cloth bags, jars and storage containers with lids, shopping for food without packaging, bartering and sewing. There are also versions derived from the letters "R" that use many words such as renewal, reduction, redesign, transformation, repair, rethinking, restoration, association, taking responsibility, rot.

In 2010, The World Organization for Early Childhood Education (OMEP) changed and developed the 7Rs at the international workshop titled "The Role of Early Childhood Education for a Sustainable Society". It is discussed as Reduce, Reuse, Respect, Reflect, Rethink, Recycle and Redistribute (Duncan, 2011). The new 7Rs are organized to include the three pillars of education for sustainable development, namely social and cultural development, economic development and environmental protection. The biggest purpose in choosing different words is to reduce waste. Many words derived from the letter "R" have been used, such as renewal, reduction, redesign, redefinition, repair, rethinking, restoration, relate, response, and rot. In short, the main purpose of choosing these words is to reduce waste.

With the cooperation established between the Ministry of Environment, Urbanization and Climate Change (CSB), Ministry of National Education (MEB) and Turkish Foundation for Combating Erosion, Afforestation and Protection of Natural Assets (TEMA) in 2018, the "Zero Waste Education Project" was launched in primary schools, and in the project, it was decided to teach zero waste to students with the "5D Rule". These are the principles created based on the principles of zero waste: *Think and Do not Consume Unless Necessary* (Düşün ve Gerekli Değilse Tüketme), *Consume Less* (Daha Az Tüket), *Evaluate and Reuse* (Değerlendir ve Yeniden Kullan), *Replace and Use for Different Purposes* (Değiştir ve Farklı Amaçla Kullan), *Recycle and Return to Nature* (Dönüştür ve Doğaya Geri Kazandır) (TEMA, 2018).

Scales related to consumption habits in the literature; Environmentally Friendly Product Consumption Behavior Scale (Karadağ Alçı *et al.*, 2023), Socially Responsible Consumption Behavior Scale (Terzi *et al.*, 2023), Recycling Awareness Scale (Ocak *et al.*, 2022), Hedonic and Utilitarian Consumption Behavior Scale (Coskun & Marangoz, 2019), Green Organizational Behavior Scale (Erbaşı, 2019), The Awareness Scale on The Recycling (Aksan & Çelikler, 2017), Conscious Consumer Scale (Buğday, 2015), Sustainable Consumption Behavior Scale (Doğan *et al.*, 2015), Environmental Responsibility Consumer Awareness Scale (Köse & Gül, 2014), Sustainable Environmental Education Attitude Scale (Afacan & Demirci-Güler, 2011). Table 1 below lists the scales, dimensions and number of items in the literature.

Existing scales have focused on psychological, socio-cultural or economic factors affecting consumption behaviors and have been addressed with the dimensions of environmental consumption behaviors, sensitivity, recycling, green consumption, anxiety, awareness, belief and attitude (Aksan & Çelikler, 2017; Buğday, 2015; Doğan *et al.*, 2015; Erbaş, 2019; Karadağ Alçı *et al.*, 2023; Karatekin, 2013; Köse & Gül, 2014; Ocak *et al.*, 2022; Tekkaya *et al.*, 2011; Terzi *et al.*, 2023). When the scales mentioned above are examined, it is seen that the steps of reuse, rejection, waste reduction and reuse, which are among the dimensions of zero waste practices, are not included. In the literature review, there is no scale measuring behavior and intention adopting the zero-waste approach. Likewise, it is thought that these concepts included in the zero-waste regulation are not sufficiently included in the curriculum and textbooks. This study was prepared based on the lack of these topics. Achieving the zero-waste goal is important in terms of supporting economic development, reviewing our consumption behaviors and reducing environmental impact. Recycling is a costly, energy-consuming and difficult process, and waste plays a major role in changing consumption behaviors by creating a negative impact on the environment.

The sample of the study consists of prospective teachers, as awareness related to environmental behavior is formed at an early age, learning at this stage tends to be more permanent, and teachers are the key figures responsible for delivering education. Today's prospective teachers are tomorrow's educators. Those who adopt nature-based and conscious consumption practices can serve as role models for their students, both personally and professionally. Teachers who evaluate, transform, and reduce their own consumption habits-while acting in accordance with environmental protection principles-can inspire similar behaviors in their students.

**Table 1.** Names, dimensions and number of items of scales and questionnaires in the national literature.

Publication Date	Name of the Scale	Sub-Dimensions	Number of Items
2023	The Environmentally Sensitive Product Consumption Behavior Scale (Karadağ Alçı <i>et al.</i> , 2023)	Environmentally friendly purchasing intention Special Norm Attitude Environmental Concern	19
2023	The Socially Responsible Consumption Behavior Scale (Terzi <i>et al.</i> , 2023)	Unidimensional	10
2022	Recycling Awareness Scale (Ocak <i>et al.</i> , 2022).	Awareness Consciousness	20
2019	The Hedonic and Utilitarian Consumption Behavior Scale (Coskun & Marangoz, 2019)	Hedonic Effect Hedonic Adaptation Passive State Impulsive Tendency Identity Mirroring	42
2019	Green Organizational Behavior Scale (Erbaş, 2019)	Environmental Sensitivity, Environmental Participation, Economic Sensitivity, Green Purchasing, Technological Sensitivity	27
2017	The Awareness Scale on The Recycling (Aksan & Çelikler, 2017)	Environmental, Educational, Economic, Administrative, Legal, Susceptibility, Media, Protection of Natural Resources, Features of Recycled Products, Biological.	48
2015	Individuals' Sustainable Consumption Behavior (Doğan <i>et al.</i> , 2015),	Environment, Unneeded Consumption, Savings, Reusability.	17
2015	Conscious Consumer Scale (Buğday, 2015)	Environmentally Conscious Consumption, Ethical Consumption, Simple Consumption, Socially Responsible Consumption	25
2014	Consumption Consciousness Depending on Environmental Responsibility Scale (Köse & Gül, 2014)	Responsibility and awareness towards the environment Consumption and Purchasing Saving	25
2013	Attitudes Towards Solid Waste and Recycling (Karatekin, 2013)	Initiative and Participation Belief Interest and Sensitivity	33
2011	Attitude Scale Towards Solid Waste and Recycling Studies (Tekkaya <i>et al.</i> , 2011)	Attitude Behavior Beliefs The Importance of Behavior Consequences Subjective Norm Perceived Expectations The Importance of Expectations Perceived Behavior Control Perceived Conditions/Situations Facilitating Conditions/Situations Behavior Intention Recycling Behavior	102



Although the principles of zero waste have been adopted as public policy, they are not yet sufficiently integrated into school curricula. Therefore, it is essential to develop a measurement tool that reveals prospective teachers' behaviors and intentions related to sustainable consumption. The scale developed in this study focuses on the key steps of the zero-waste approach in order to evaluate prospective teachers' consumption habits and intentions. The sub-dimensions of the scale-namely waste reduction, critically examining the need for consumption, rejecting unnecessary items, reusing, recycling, and composting-are structured to offer a new perspective to the field. Moreover, the scale is designed to measure both behavior and intention, with the aim of illuminating possible discrepancies between the two and offering insights for future research.

Numerous theoretical models have been developed to understand individuals' attitudes and behaviors toward environmental issues. These models indicate that knowledge alone is insufficient to foster environmental behavior; instead, values, attitudes, social norms, perceived personal responsibility, and identity development all play critical roles (Ajzen, 1991; Gifford & Nilsson, 2014). The Theory of Planned Behavior emphasizes the cognitive structures underlying behavioral intentions, while Kollmuss and Agyeman (2002) highlight the "value-action gap," whereby knowledge and values do not necessarily lead to behavior. In this context, prospective teachers need a holistic educational approach that integrates cognitive, affective, and experiential learning in order to develop an environmentally responsible identity. Additionally, Steg and Vlek (2009) argue that contextual and structural factors, beyond individual motivation, significantly influence pro-environmental behavior. They emphasize that nature-based learning environments have transformative potential for the development of ecological identity. In line with these perspectives, this study aims to evaluate the consumption behaviors and intentions of prospective teachers using a zero-waste framework enriched through nature-based environmental education.

## 2. METHOD

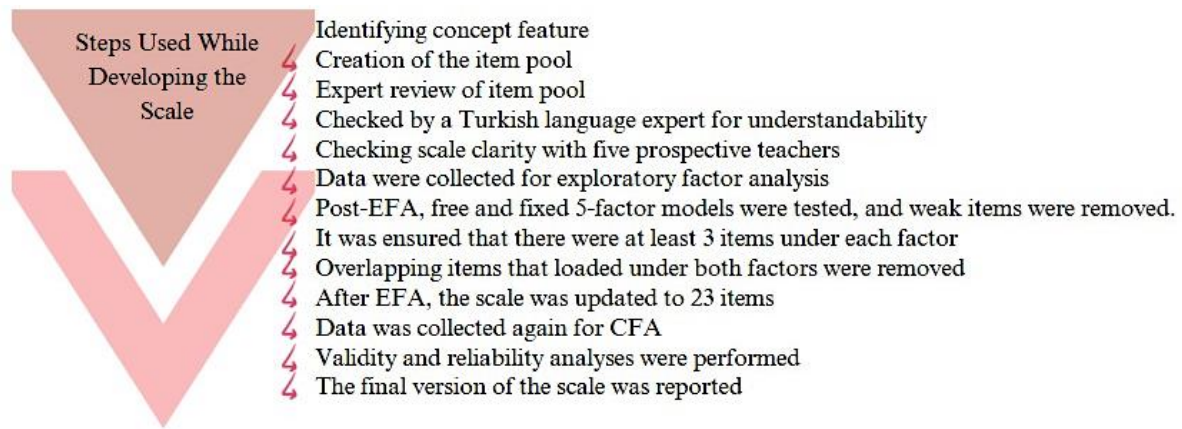
In this section, the study participants, the "Sustainable Consumption Behavior and Intention Scale" development process and data analysis techniques are presented

### 2.1. Participants

Convenience sampling is a type of sampling in which the researcher applies a survey to the people he can most easily reach. It is related to the fact that it is easier to include the individuals or groups to be researched in the research process or to reach them (Yıldırım & Şimşek, 2008). The study was conducted within the Faculty of Education at a public university located in the city center of Denizli. In order for the scale to represent the target population, efforts were made to reach pre-service teachers enrolled in different teacher education programs, and data were collected from participants at various grade levels. In order to obtain valid and reliable results, it was aimed for the sample to represent the universe as accurately as possible. For this reason, data were collected in general culture courses where students from 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup> grades could take courses. These courses were selected both because the number of students was high and because students from each department took these courses. While the data collected for the exploratory factor analysis were collected from teacher candidates who took the elective vocational knowledge course in the 2022-2023 fall semester ( $n = 242$ ), it was collected in the 2022-2023 spring semester ( $n = 280$ ) for the confirmatory factor analysis.

### 2.2. Scale Development Process

The basic stages in the process of creating an original scale show similar characteristics. These stages can be grouped under the name of examining the theoretical structure and creating items, consulting expert opinion, pilot application, main application, and validity and reliability analyses (DeVellis, 2014; Hinkin, 1998; Karasar, 2014; Seçer, 2015; Şeker & Gençdoğan, 2014). In this study, the steps in [Figure 3](#) were followed:



**Figure 3.** Steps followed while developing the scale.

While preparing the scale items, 3R, 5R, 7R and 5D, literature review, semi-structured interviews with teacher candidates and interviews with field experts were used. While scanning the literature, Reject, Reduce, Reuse, Recycle and Rot, which Johnson (2019) covers in the zero waste process stated in his book; Think, transform, change, consume less and evaluate (TEMA, 2018) and 5D for the sustainability of nature: "Think, Do Not Consume Unless Necessary", "Consume Less", "Evaluate, Reuse", "Change, Different" The steps "Use for Purpose" and "Transform, Let Nature Win" (TEMA, 2019) are the concepts taken into consideration when creating the sub-dimensions of the scale. While preparing the consumption behaviors scale, 59 items were included in light of the classifications specified. 10-12 items were written under each sub-dimension. In order to avoid any misinterpretation in the items, no reverse items were written.

The scale consists of two parts. In the first part, they were asked to rate how they behave in daily life (Column A), while in the second part, they were asked how they wanted to behave and their intentions (Column B). It was stated that they had to answer in both columns while filling out the scale. Each item was rated as Never (1), Rarely (2), Sometimes (3), Often (4) and Always (5). The data to be used for exploratory factor analysis were collected from 242 teacher candidates studying at Pamukkale University Faculty of Education in the 2022-2023 fall semester. After EFA, the analyses were carried out by both freeing the number of factors and fixing them to 5 dimensions. Considering the item difficulty index and factor loadings, items that loaded below .30 and items that loaded on more than one factor were removed. There are at least 3 items under each factor. Finally, the scale was organized into 5 factors and 23 items.

### 2.3. Expert Opinion Consultancy

While consulting an expert opinion in the study, the scale items were written under each sub-dimension according to the literature. It was collected from four Science Education field experts in forms consisting of suitable, slightly corrected, correctable, not suitable and suggestions for each item. Later, the items were discussed in an online meeting with three field experts and two more items were added to the item pool, creating the first version of the scale consisting of 59 items. In addition, for this version of the scale, the opinion of a Turkish teacher was consulted and edited in terms of language control.

### 2.4. Analysis of Data

Before proceeding with the analysis of the data obtained in the study, frequency analysis was used to check whether there were any incorrectly entered data, and when deemed necessary, the participants' scales were re-checked and edited. After the data set was edited and checked, the number of missing data values was checked. According to Güzeller (2016), if there are missing values in a data set, these values can be deleted or different assignment methods can be used. Since there were 0 to 8 missing data under each item and the missing data rate was less

than 5%, it was preferred to delete the missing values in this study. In addition, since Little's MCAR test result accepted the  $H_0$  hypothesis, the missing data were deleted because the data loss was random and independent of other independent variables ( $p = .138$ ).

In order to ensure the construct validity of the scale, preliminary analyses were conducted prior to factor analysis. One of the fundamental criteria in determining the suitability of the data for factor analysis is whether a sufficient sample size has been reached (Munro, 2005). In this context, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity were applied to evaluate the adequacy of the data. First, the KMO value was calculated to assess whether the sample size and the correlations among variables were appropriate for factor analysis. According to both Field (2018) and Büyüköztürk (2002), a KMO value above 0.80 is considered meritorious, indicating that the sample is sufficient to uncover latent structures through factor analysis. The KMO statistic reflects the proportion of common variance among variables, which is essential for producing stable and interpretable factor solutions. In addition, Bartlett's Test of Sphericity was employed to test the null hypothesis that the correlation matrix is an identity matrix, implying no significant relationships among the variables. A statistically significant result ( $p < .05$ ) indicates that there are adequate inter-item correlations to proceed with factor analysis (Büyüköztürk, 2002; Field, 2018). Although Bartlett's test does not directly assess validity, it provides important preliminary evidence regarding the appropriateness of factor analysis, thereby contributing to construct validation.

The statistically significant results obtained from both the KMO and Bartlett tests confirmed that the dataset was suitable for factor analysis and supported the structural validity of the developed scale. Multivariate normality assumption calculations were performed for both data sets. In the first data set where EFA was performed, 36 items were removed, and it was revealed that multivariate normality was not achieved ( $p < .05$ ) for 23 items. Likewise, in the second data set where CFA was performed, it was determined that multivariate normality was not achieved ( $p < .05$ ).

When selecting items from the item pool, items that did not provide sufficient value in any dimension were removed and the analyses were re-run. Items that gave the best eigenvalues under factor distribution with different combinations were selected. In this process, items that did not load significantly on the dimensions that were expected to load at the beginning of the scale were also removed. In addition, items with negative values were removed even though they were not written as reverse items. Items that could be included in the 5 determined dimensions were included in the scale, and thus the number of scale items was determined as 23. The 36 items were removed because their factor load values were overlapping and they were not included in any dimension.

As a result of the analysis, it was determined that the correlation coefficients between the factors varied between 0.153 and 0.389. Based on this information, it can be said that there is a weak level of relationship between the factors (Durmuş *et al.*, 2013; Kalaycı, 2014; Şencan, 2005). For this reason, Oblimin oblique rotation factor rotation, which assumes that there is a relationship between the factors in EFA, was adopted (Büyüköztürk, 2002b). Principal Axis Factoring was used in this study. It is used in scale development studies to discover latent factors and to reveal the explanatory power of these factors for data (Büyüköztürk, 2002b).

Within the scope of EFA, Principal Axis Factoring, Kaiser-Meyer-Olkin (KMO) test and Bartlett sphericity test, percentage of explained variance calculation and calculation of factor load values were used (DeVellis, 2014). To determine the number of factors, Horn (1965)'s parallel analysis method, line graph and eigenvalues were examined. The number of factors was freed at the beginning of the analysis. Since the factor was more than the theoretical framework and two items were found under the factors, the number of factors was fixed and continued. This was supported in the parallel analysis method.

In this study, assumptions were tested before the validity and reliability analyses of the scale development. For multivariate normality, *p-values* of both skewness and kurtosis statistics are expected to be greater than 0.05. Mardia coefficient was found to be significant when calculated with the help of Jamovi Version 2.4.14. EFA and CFA were performed with the same application. In order to determine the reliability of the factors, both Cronbach's alpha ( $\alpha$ ) and McDonald's omega ( $\omega$ ) coefficients were calculated (R Core Team, 2022; Revelle, 2023; Rosseel, 2023; The Jamovi Project, 2023). Cronbach's alpha is a widely used reliability measure, but it assumes tau-equivalence (i.e., that all items have equal factor loadings). In contrast, McDonald's omega offers a more robust estimate of internal consistency, particularly when item loadings vary across the scale (Dunn *et al.*, 2014). Table 2 below presents the Cronbach's alpha ( $\alpha$ ) and McDonald's omega ( $\omega$ ) reliability coefficients calculated for both the EFA and CFA samples, demonstrating the internal consistency of the Sustainable Consumption Behavior and Intention Scales. In this study, confirmatory factor analysis (CFA) was conducted, and the Diagonally Weighted Least Squares (DWLS) estimation method was used. The DWLS method is preferred when working with categorical (ordinal) data as it provides more reliable parameter estimates (Li, 2016). It is a specific type of Weighted Least Squares (WLS) estimation that utilizes only the diagonal elements of the covariance matrix. This method is known for its robustness against violations of normality assumptions.

**Table 2.** Consumption habits behavior and intention scale reliability value.

	Type of Scale	<i>n</i>	$\alpha$	$\omega$
EFA	Consumption Behavior Scale	242	.870	.872
	Consumption Intention Scale		.910	.925
CFA	Consumption Behavior Scale	280	.846	.850
	Consumption Intention Scale		.907	.911

Note. Diagonally Weighted Least Squares (DWLS) estimation method was used.

### 3. RESULTS

This section contains findings regarding the validity and reliability studies of the scale.

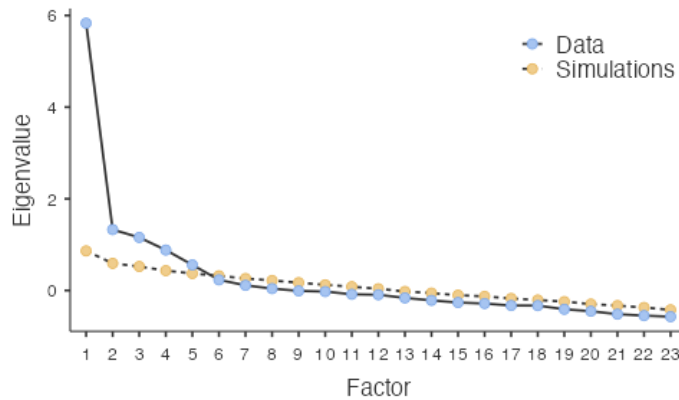
#### 3.1. Findings Regarding Validity

The suitability of the data set obtained from 242 participants for exploratory factor analysis (EFA) was assessed using the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity. The KMO value was found to be .826, indicating that the sample size and inter-item correlations were sufficient for conducting factor analysis. According to established benchmarks, KMO values between .80 and .89 are interpreted as "great" in terms of sampling adequacy (Field, 2018). Moreover, as the KMO coefficient approaches 1, it reflects a stronger degree of shared variance among variables, with 1.00 representing perfect adequacy (Field, 2018). Bartlett's Test of Sphericity yielded a statistically significant result,  $\chi^2(253) = 1801$ ,  $p < .001$ , indicating that the correlation matrix is not an identity matrix and that correlations among items were sufficiently large to justify factor analysis (Büyüköztürk, 2002; Tabachnick & Fidell, 2007). These findings collectively support the suitability of the data for factor analysis and strengthen the construct validity of the scale.

In the parallel analysis method used to discover how many dimensions the scale has it was seen that the number of factors was 5. In the line chart in Figure 4, it can be seen that the fifth factor is where the data and suggested lines intersect. The five-factor structure of the scale is shown in Table 3, considering its eigenvalue within the framework of the literature and theory, as its eigenvalue is above 1. To determine the number of factors, both statistical outputs and theoretical underpinnings were considered. Initially, the scree plot and eigenvalue criteria (greater than 1) suggested a five-factor solution. This decision was further supported by the theoretical framework of sustainable consumption behavior and intention, which is grounded



in Ajzen's (1991) Theory of Planned Behavior and extended by Kollmuss and Agyeman's (2002) model of pro-environmental behavior.



**Figure 4.** Parallel analysis line graph.

The five-factor structure aligns with the multidimensional nature of sustainable consumption, encompassing components such as awareness, intention, behavior, responsibility, and ethical concern, as suggested in previous literature (Gifford & Nilsson, 2014; Steg & Vlek, 2009). Thus, the factor structure reflects not only empirical findings but also established theoretical constructs in the field. Table 3 also includes the percentages of variance explained as a result of the analysis.

**Table 3.** EFA results.

Factor	SS Loadings	% of Variance	Cumulative %
1	2.64	11.48	11.5
2	2.33	10.13	21.6
3	2.28	9.90	31.5
4	1.79	7.80	39.3
5	1.79	7.76	47.1

When Table 3 above is examined, it can be seen that the number of factors is compatible with 5. In the light of the theoretical structure, parallel analysis method and eigenvalues, the suitability of the 5-factor structure was decided. In addition, in the percentage of variance explained by EFA, 47.1% of the total variance was explained and is shown in the table.

The factor loading values and originality of the version of the scale consisting of 23 items are shown in Appendix A. The dimensions of the scale were labeled as Reconsider, Reduce, Reuse, Recycle, and Rot based on the related literature. The factor loading values of the scale items are also presented in Appendix A. When it is examined, it is seen that the factor loadings of the scale items range between .358 and .814. Since all values are greater than the critical cut-off points of .30, the items were retained in the scale (Büyüköztürk, 2002; Tabachnick & Fidell, 2007), the items can be included under the factor structure. In EFA, it is desirable that the items have a high relationship with the factors, and it can be said that the scale shows a distribution consistent with the characteristics to be measured.

CFA was conducted to verify the 5 factor and 23 item structure determined in EFA on the second data set of 280 people in the study and to reveal the causality relationship between the determined items. Modification suggestions were also evaluated from the analysis results. Modifications were made between items 4 and 14, 9 and 23, 5 and 10, 15 and 18. It was thought that the error variances might be related because the items in question were of the same size and the statements were similar to each other and expressed the same behavioral tendency. It was determined that there was an improvement in the fit indices with the modification made.



As a result of CFA, it was determined that all items explained the main variance at a statistically significant level ( $p < .001$ ).

In this study, confirmatory factor analysis (CFA) was conducted, and the Diagonally Weighted Least Squares (DWLS) estimation method was used. The DWLS method is preferred when working with categorical (ordinal) data as it provides more reliable parameter estimates (Li, 2016). It is a specific type of Weighted Least Squares (WLS) estimation that utilizes only the diagonal elements of the covariance matrix. This method is known for its robustness against violations of normality assumptions.

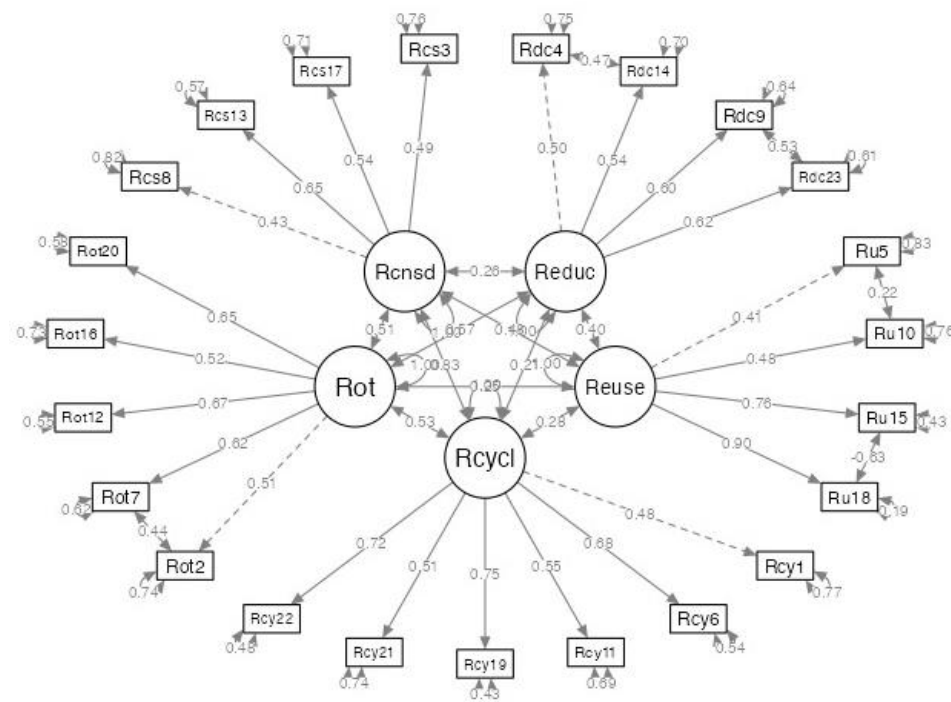
When looking at the other fit indices in Table 4, it is noteworthy that all values except for the NFI value show an acceptable or good fit. It is generally recommended in the literature to evaluate model fit using multiple indices rather than relying on a single one, as each index provides different information about model performance (Cabrera-Nguyen, 2010; Schermelleh-Engel *et al.*, 2003). The RMSEA value was .050, with a 95% confidence interval of [.043, .065], indicating an acceptable model fit (MacCallum *et al.*, 1996). Based on these fit indices, it can be concluded that the model exhibits a good fit to the data and adequately reflects the underlying construct. In other words, the CFA results confirm the proposed structure of the Sustainable Consumption Behavior and Intention Scale (SuCBIS), consisting of 23 items and 5 factors.

**Table 4.** CFA fit indices and values.

Fit Index	Analysis Value	Shown Fit
$\chi^2 / df$	365 / 216 = 1.68	Good Fit
RMSEA	.050	Good Fit
SRMR	.067	Acceptable Fit
CFI	.979	Good Fit
NFI	.951	Good Fit
NNFI	.976	Good Fit

The confirmatory factor analysis (CFA) results indicated that all factor loadings were statistically significant ( $p < .001$ ), with standardized estimates ranging from .455 to .900. Standard errors were within acceptable ranges, and all z-values exceeded the critical threshold for statistical significance. These findings confirm the adequacy of the measurement model and demonstrate that the items reliably reflect their respective latent constructs. The path diagram derived from the CFA is presented in Figure 5.

In order to evaluate the reliability of the scale, both Cronbach's alpha and McDonald's omega coefficients were calculated. Although Cronbach's alpha is widely used in the social sciences, it relies on certain assumptions such as equal item means which are often violated in practice. When these assumptions are not met, alpha may underestimate the actual reliability of the instrument. Therefore, McDonald's omega was also computed in this study, as it provides a more accurate estimate of internal consistency by taking into account the factor loadings across items. The reliability coefficients for the scale dimensions based on the 5R framework are presented in Table 5. For the dimensions of Rot, Recycle, Reuse, and Reduce, both Cronbach's alpha and McDonald's omega values exceeded the recommended threshold of .70, indicating satisfactory internal consistency. However, for the Reconsider dimension, the reliability scores were relatively low ( $\alpha = .616$ ,  $\omega = .560$ ), suggesting potential issues with the construct consistency of this dimension. Although the Reconsider dimension demonstrated a reliability coefficient of  $\alpha = .616$ , which is considered acceptable according to some criteria (e.g., Nunnally & Bernstein, 1994), this value still falls below the commonly preferred threshold of .70. Therefore, these findings suggest that the items within the Reconsider factor may benefit from further revision or refinement to improve internal consistency and better represent the underlying construct.



**Figure 5.** Path diagram of the model.

**Table 5.** Reliability coefficients of scale dimensions.

Dimensions	$\alpha$	$\omega$
Rot	.811	.773
Recycle	.808	.782
Reuse	.746	.713
Reconsider	.616	.560
Reduce	.803	.714

#### 4. DISCUSSION and CONCLUSION

In this study, the Sustainable Consumption Behavior and Intention Scale (SuCBIS) was developed to assess both behavioral and intentional dimensions of sustainable consumption among prospective teachers. Since no existing scale in the literature evaluates these two constructs simultaneously, SuCBIS stands out as an original and significant contribution. The validity and reliability analyses confirmed that the scale demonstrates high internal consistency both at the total and subscale levels, supporting its potential for use in future applications.

However, the findings should be interpreted in light of certain limitations. Although the five-factor structure of the scale was supported by first-order confirmatory factor analysis (CFA), a second-order CFA or bifactor model was not conducted to determine whether a meaningful total score could be calculated. This constitutes a methodological limitation. Future research should consider testing second-order or bifactor models to provide more robust evidence regarding the unidimensionality of the scale. Until such evidence is available, it is recommended that subscale scores be reported individually rather than relying solely on a total score.

The variance explained by the scale was 47%, which falls within the generally accepted range of 40%–60% for similar instruments (Atabek-Yiğit *et al.*, 2020; Gökçe *et al.*, 2024). However, to enhance the explanatory power of the scale, future studies may revise existing items or develop new ones with larger and more diverse samples. The relatively modest variance explained may be attributed to the multifaceted nature of sustainable consumption behavior,

which encompasses cognitive, affective, social, and contextual dimensions. Similar to findings in prior research, it is not uncommon for scales measuring complex behaviors such as pro-environmental or sustainable actions to report explained variance below 50% (Gifford & Nilsson, 2014; Kollmuss & Agyeman, 2002). Additionally, since the current study was conducted exclusively with prospective teachers, its generalizability to other professional or demographic groups is limited. Testing the scale across various populations would expand its applicability and relevance.

This scale measures not only what individuals intend to do but also what they actually do, making it possible to examine the alignment or mismatch between intention and behavior. In complex constructs such as sustainable consumption which include cognitive, affective, and behavioral dimensions, identifying discrepancies between intention and behavior is critical for informing educational interventions. In this study, the grouping of "Reuse" and "Reconsider" items under the "Recycle" factor and the removal of several items from the behavioral dimension due to low factor loadings may indicate conceptual limitations in participants' understanding of zero-waste principles.

This finding is consistent with previous literature. Harman and Yenikalaycı (2019) found that pre-service teachers had limited awareness of the zero-waste approach, often interpreting it primarily through the lens of waste management and recycling, with minimal reference to reuse. Similarly, national studies tend to focus more on recycling rather than the broader zero-waste framework (Bulut & Çavuldur, 2017; Mutlu, 2013; Ural-Keleş & Keleş, 2018). This result is also in line with recent international literature emphasizing the need to address the gap between sustainability-related intentions and actual behaviors (Fischer *et al.*, 2017). For instance, Liu *et al.* (2012) found that even among individuals with positive attitudes toward green consumption, behavioral change often remained limited without adequate educational intervention. These findings underline the global relevance of the SuCBIS, which not only assesses both behavioral and intentional dimensions but also offers a structured tool for identifying educational needs in sustainable consumption practices across various contexts. Thus, SuCBIS serves as an important tool for reminding educators and learners that zero waste involves more than recycling, emphasizing the importance of reducing consumption, reusing materials, and rethinking habits.

These findings also support theoretical frameworks that distinguish between behavioral intention and actual behavior. According to Ajzen's (1991) Theory of Planned Behavior, behavioral intentions are shaped by attitudes, subjective norms, and perceived behavioral control, but do not always translate into action—a phenomenon also described as the "intention-behavior gap" by Kollmuss and Agyeman (2002). In this sense, SuCBIS not only captures this gap but also contributes a behaviorally grounded, context-specific measurement tool for sustainability education. Furthermore, the integration of zero-waste principles—such as reject, reduce, and compost—into the subdimensions of the scale expands the field beyond abstract environmental attitudes and toward actionable, measurable change. In addition, by highlighting the divergence between "Reconsider" and "Reuse," this study contributes to the growing body of research on ecological identity formation, which emphasizes experiential and values-based learning for sustainable behavior (Clayton, 2003; Steg & Vlek, 2009).

SuCBIS enables the evaluation of individuals' tendencies toward sustainable consumption at both behavioral and intentional levels and allows for the identification of intention-behavior discrepancies, which can inform both educational strategies and policy design. It is critical that sustainability education moves beyond recycling and incorporates a more holistic approach grounded in the principles of reducing, reusing, and rethinking. The developed scale can be employed as an assessment tool in teacher training programs, environmental education practices, and school-based sustainability initiatives.

While the present study provides valuable insights into sustainable consumption behavior and intentions through the development of the SuCBIS, several limitations should be

acknowledged. First, although the five-factor structure of the scale was supported by first-order confirmatory factor analysis (CFA), no second-order or bifactor model was conducted to assess the viability of a meaningful total score. Therefore, it is recommended that subscale scores be interpreted separately, and future studies should explore higher-order models to further validate the structure. Second, the sample consisted exclusively of prospective teachers from specific departments in a single region of Türkiye. This limits the generalizability of the findings to other populations, educational systems, or cultural contexts. Third, the study relied on self-reported data, which may be affected by social desirability bias or inaccurate self-assessment. Additionally, although the internal consistency of the scale was high, longitudinal data were not collected; thus, test–retest reliability and temporal stability remain unexamined. Finally, the cross-sectional design restricts the ability to observe changes in sustainable consumption behavior and intentions over time. These limitations may, in part, reflect the multifaceted and complex nature of sustainable consumption, which is influenced by a broad range of psychological, contextual, and socio-cultural factors (Gifford & Nilsson, 2014; Kollmuss & Agyeman, 2002). Future research should aim to address these limitations by employing longitudinal and cross-cultural designs with more diverse samples to enhance the scale’s robustness and applicability in broader educational and societal contexts.

In conclusion, SuCBIS enables the evaluation of individuals’ tendencies toward sustainable consumption at both behavioral and intentional levels and allows for the identification of intention–behavior discrepancies, which can inform both educational strategies and policy design. It is critical that sustainability education moves beyond recycling and incorporates a more holistic approach grounded in the principles of reducing, reusing, and rethinking. The developed scale can be employed as an assessment tool in teacher training programs, environmental education practices, and school-based sustainability initiatives. Moreover, its use can support policymakers and curriculum developers by providing data-driven insights into the effectiveness of sustainability education and guiding the design of interventions that promote long-term behavior change. Practically, the scale offers a framework for educators to diagnose and improve prospective teachers’ readiness to foster sustainable habits in their future classrooms. It can be used to tailor professional development programs, integrate sustainability more explicitly into teacher education curricula, and evaluate the impact of pedagogical strategies aimed at promoting eco-responsible behavior. Thus, SuCBIS not only contributes theoretically but also responds to practical educational needs in the field of environmental and sustainability education.

### Acknowledgments

This research is part of the thesis titled “The Impact of Nature-Based Environmental Education on Pre-Service Teachers’ Environmental Identity and Pedagogical Competence”.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Pamukkale University Ethics Committee, 68282350/2022/G02.

### Contribution of Authors

**Merve Eker Çelebi:** Conceptualization, Data collection, Analysis, and Drafting the manuscript. **Fatma Taşkın Ekici:** Supervision, Methodological support, and Critical revision of the manuscript.

### Orcid

Merve Eker Çelebi  <https://orcid.org/0000-0002-8805-6214>

Fatma Taşkın Ekici  <https://orcid.org/0000-0001-7798-6021>



## REFERENCES

- Afacan, Ö., & Demirci-Güler, M.P. (2012). Development of attitude scale in the context of sustainable environmental education. *Energy Education Science and Technology Part B: Social and Educational Studies*, 4(4), 2479-2488.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Aksan, Z., & Çelikler, D. (2017). The development of a recycling awareness scale for prospective science teachers. *Educational Studies*, 43(5), 567-583. <https://doi.org/10.1080/03055698.2017.1312289>
- Atabek-Yiğit, E., Yavuz-Topaloğlu, M., & Balkan-Kıyıcı, F. (2020). Recycling Scale for secondary school students: Scale development and reliability. *PESA International Journal of Social Studies*, 6(3), 244-254. <https://doi.org/10.25272/j.2149-8385.2020.6.3.04>
- Bektaş, Ş. (2020, March 2). *Sıfır atık [Zero-waste]* [PowerPoint slides]. Çevre ve Şehircilik Bakanlığı, Çevre Yönetimi Genel Müdürlüğü [Ministry of Environment and Urbanization, General Directorate of Environmental Management]. <https://pagev.org/upload/files/1%20%C5%9EULE%20BEKTA%C5%9E.pptx>
- Buğday, E.B. (2015). *Bilinçli tüketici ölçeği geliştirme çalışması [Development of the Conscious Consumer Scale]* [Doctoral dissertation]. Hacettepe University.
- Bulut, E., & Çavuldur, L. (2017). The impact of recycled paper pulp usage as a creative material in the visual arts education in developing an awareness and habit for paper recycling among the students. *International Journal of Afro-Eurasian Research*, 2(4), 187-208. <https://dergi.park.org.tr/en/download/article-file/385029>
- Büyüköztürk, Ş. (2002). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum* [Handbook of data analysis for the social sciences: Statistics, research design, SPSS applications, and interpretation (1<sup>st</sup> Edition)]. Pegem Publishing.
- Büyüköztürk, Ş. (2002b). Factor analysis: basic concepts and its use in scale development. *Educational Administration in Theory & Practice*, 32, 470-483.
- Büyüköztürk, Ş., Kılıç-Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2014). *Bilimsel araştırma yöntemleri* [Scientific research methods]. (16<sup>th</sup> Edition). Pegem Academy.
- Cabrera-Nguyen, E.P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, 1(2), 99-103. <https://doi.org/10.5243/jsswr.2010.8>
- Clayton, S. (2003). Environmental identity: A conceptual and an operational definition. In S. Clayton & S. Opatow (Eds.), *Identity and the natural environment: The psychological Significance of Nature* (pp. 45-65). MIT Press.
- Coskun, T., & Marangoz, M. (2019). Development of the hedonic and utilitarian consumption behavior scale: reliability and validity study. *Business and Economics Research Journal*, 10(2), 517-540. <https://doi.org/10.20409/berj.2019.183>
- DeVellis, R.F. (2014). *Scale development: Theory and applications* (4<sup>th</sup> ed.). Sage.
- Doğan, O., Bulut, Z.A., & Kökalan-Çımrın, F. (2015). A scale development study to measure individuals' sustainable consumption behavior. *Trends in Business and Economics*, 29(4), 659-678.
- Duncan, E. (2011). *Report Part 2 – ESD in practice*. OMEP (Organisation Mondiale Pour L'Éducation Préscolaire).
- Dunn, T.J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399-412. <https://doi.org/10.1111/bjop.12046>
- Durmuş, B., Yurtkoru, E.S., & Çinko, M. (2013). *Sosyal bilimlerde SPSS'le veri analizi* [Data analysis with SPSS in the social sciences] (5<sup>th</sup> ed.). Beta Yayıncılık.



- Erbaşı, A. (2019). Yeşil Örgütsel Davranış Ölçeği: Bir ölçek geliştirme çalışması [Green Organizational Behavior Scale: A scale development study]. *Istanbul Management Journal*, 86, 1-23. <http://doi.org/10.26650/imj.2019.86.0001>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5<sup>th</sup> ed.). SAGE Publications.
- Gifford, R., & Nilsson, A. (2014). Personal and social factors that influence pro-environmental concern and behaviour: A review. *International Journal of Psychology*, 49(3), 141-157. <https://doi.org/10.1002/ijop.12034>
- Gökçe, N., Çetinkaya-Aydoğdu, C., Arslan, E., Bayram, F.Ö., & Akbaş, A. (2024). Ortaokul öğrencilerine yönelik yeniden kazanım tutum ölçeğinin geliştirilmesi. [Development of a recycling attitude scale for middle school students]. *Kırşehir Eğitim Fakültesi Dergisi*, 25(1), 784-816.
- Güzeller, C.O. (2016). *Herkes için çok değişkenli istatistik* (1. Baskı) [Multivariate statistics for everyone (1<sup>st</sup> ed.)]. Maya Akademi.
- Harman, G., & Yenikalaycı, N. (2020). Awareness of preservice science teachers on zero waste approach. *Pamukkale University Journal of Education*, 50, 138-161. <https://doi.org/10.977/9/pauefd.589781>
- Hinkin, T.R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1(1), 104-121. <https://doi.org/10.1177/109442819800100106>
- Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179-185. <https://doi.org/10.1007/BF02289447>
- Johnson, B. (2019). *Sıfır atık ev* [Zero waste home] (Ö. Kocaefe, Trans.). Sinek Sekiz Yayınları. (Original work published 2013)
- Kalaycı, Ş. (Ed.). (2014). *SPSS uygulamalı çok değişkenli istatistik teknikleri* (6. baskı) [Multivariate statistical techniques with SPSS applications (6<sup>th</sup> ed.)]. Asil Yayın Dağıtım.
- Karatekin, K. (2013). Developing a scale to measure pre-service teachers' attitudes towards solid waste and recycling: a validity and reliability study. *International Journal of Eurasia Social Sciences*, 4 (10), 71-90.
- Karadağ-Alçı Ş., Geçkil T., & Aksu S. (2023). Turkish adaptation of the environmentally sensitive product consumption behavior scale: A validity and reliability study. *Gümüşhane University Journal of Social Sciences Institute*, 14(2), 540-553. <https://doi.org/10.36362/gumus.1224548>
- Karasar, N. (2014). *Bilimsel araştırma yöntemi* [Scientific research method] (26<sup>th</sup> ed.). Nobel Yayıncılık.
- Kollmuss, A., & Agyeman, J. (2002). Mind the gap: Why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research*, 8(3), 239-260. <https://doi.org/10.1080/13504620220145401>
- Köse, E.Ö., & Gül, Ş. (2014). Öğretmen adayları için çevre sorumluluğuna bağlı tüketim bilinci ölçeğinin geliştirilmesi [Development of consumption consciousness depending on environmental responsibility scale for prospective teachers]. *Journal of Educational Sciences & Practices*, 13, 257-277.
- Li, C.H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936-949. <https://doi.org/10.3758/s13428-015-0619-7>
- Liu, X., Wang, C., Shishime, T., & Fujitsuka, T. (2012). Sustainable consumption: Green purchasing behaviours of urban residents in China. *Sustainable Development*, 20(4), 293-308. <https://doi.org/10.1002/sd.484>
- MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130-149. <https://doi.org/10.1037/1082-989X.1.2.130>
- Munro, B.H. (2005). *Statistical methods for health care research* (5<sup>th</sup> ed.). Lippincott Williams & Wilkins.

- Mutlu, M. (2013). "Recycling" concepts perceptions of grade eight students: phenomenographic analysis. *Anthropologist* 16(3), 663-669.
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3<sup>rd</sup> ed.). McGraw-Hill.
- Ocak, G., Olur, B., & Aydın, T.Y. (2022). Recycling awareness scale: A scale development study. *African Educational Research Journal*, 10(2), 107-116. <https://doi.org/10.30918/AE RJ.102.22.009>
- R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (R package). <https://doi.org/10.32614/CRAN.package.psych>
- Rosseel, Y. (2023). *lavaan: An R Package for Structural Equation Modeling*. *Journal of Statistical Software*, 48(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.
- Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci: SPSS ve LISREL uygulamaları* [The process of developing and adapting psychological tests: SPSS and LISREL applications]. Anı Yayıncılık.
- Şeker, H., & Gençdoğan, B. (2014). *Psikolojide ve eğitimde ölçme aracı geliştirme* [Developing measurement tools in psychology and education]. Nobel Yayıncılık.
- Song, Q., Li, J., & Zeng, X. (2015) Minimizing the increasing solid waste through zero waste strategy. *Journal of Cleaner Production*, 104, 199-210. <https://doi.org/10.1016/j.jclepro.2014.08.027>
- Steg, L., & Vlek, C. (2009). Encouraging pro-environmental behaviour: An integrative review and research agenda. *Journal of Environmental Psychology*, 29(3), 309-317. <https://doi.org/10.1016/j.jenvp.2008.10.004>
- Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik* [Reliability and validity in social and behavioral measurements]. Seçkin Yayıncılık.
- Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Pearson.
- Tekkaya, C., Kılıç, D.S., & Şahin, E. (2011, April 27–29). *Geridönüşüm davranışının planlanmış davranış teorisi ile açıklanması: Sürdürülebilir bir kampüs için geri dönüşüm anketi* [Explaining recycling behavior with the theory of planned behavior: Recycling survey for a sustainable campus] [Paper presentation]. 2nd International Conference on New Trends in Education and Their Implications (ICONTE), Antalya, Türkiye.
- TEMA [Turkish Foundation for Combating Erosion, Afforestation and Protection of Natural Assets]. (2018, December 26). Sıfır Atık on binlerce çocuğa ulaştırılacak [Zero waste will be delivered to tens of thousands of children] [Press release]. <https://www.tema.org.tr/basin-odasi/basin-bultenleri/sifir-atik-on-binlerce-cocuga-ulaştirilacak>
- TEMA. (2019). *Doğanın sürdürülebilirliği için 5D* [5D for the sustainability of nature]. Turkish Foundation for Combating Erosion, Afforestation and Protection of Natural Assets. <https://sifiratiktema.org/?/ortaokul>
- Terzi, H., Baydar, V., Tosun, E.K., Sayın, M.E., & Ok, Ş. (2023). Sosyal sorumlu tüketim davranışı ölçeğinin kısaltılmış Türkçe versiyonunun geçerlilik ve güvenilirlik çalışması [The Turkish Short Version of The Socially Responsible Consumption Behavior Scale (SRCBS): A Scale Adaptation]. *Studies on Social Science*, 3(2), 83-102. <http://dx.doi.org/10.53035/S OSSCI.67>
- The jamovi project (2023). *Jamovi* (Version 2.4) [Computer Software]. <https://www.jamovi.org>
- Ural-Keleş, P., & Keleş, M.İ. (2018). İlkokul 3. ve 4. sınıf öğrencilerinin geri dönüşüm kavramı ile ilgili algıları [Perceptions of the 3rd and 4th Grade Students of Elementary School about The Concept of "Recycling"]. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi*, 20(2), 481-498. <https://doi.org/10.17556/erziefd.404816>

- Yıldırım, A., & Şimşek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri* [Qualitative research methods in the social sciences] (7<sup>th</sup> ed.). Seçkin Yayıncılık.
- Zaman, A. (2015). A comprehensive review of the development of zero waste management: Lessons learned and guidelines. *Journal of Cleaner Production*, 91, 12-25. <https://doi.org/10.1016/j.jclepro.2014.12.013>
- Zero Waste Regulation. (2019, July 12). *Resmi Gazete*, (30829). <https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=32659&MevzuatTur=7&MevzuatTertip=5>

## APPENDIX

Appendix A presents the factor loading values obtained from the exploratory factor analysis (EFA) conducted using the principal axis factoring method. As shown in the table, the items loaded clearly onto their respective factors, supporting the factorial validity of the scale. These results are consistent with the findings reported in the main text (see Section 3.1).

**Appendix A.** EFA factor loading values according to principal axis factoring - Turkish version.

Item	Sub-Dimensions					Uniqueness
	Rot (Kompost Yap)	Recycle (Geri Dönüştür)	Reuse (Yeniden Kullan)	Reconsider (Düşün)	Reduce (Azalt)	
7. Gıda atıklarını kompost yaparak zengin içerikli toprağa/gübreye dönüştürebilirim.	0.814					0.296
12. Farklı kompost türleri yaşımda kullanırım.	0.649					0.417
16. Kompost türlerine göre atıklarımı ayrıştırabilirim.	0.629					0.506
20. Organik atıklarımın soğuk/sıcak kompost yapabilirim.	0.601					0.520
2. Organik (doğada çözünebilir) atıklarımı çürütebilirim.	0.542					0.635
19. Satın aldığım ürünlerin geri dönüştürülebilir olmasına dikkat ederim.		0.707				0.479
1. Geri dönüşüm sürecinin maliyetli bir süreç olduğu için daha az atık oluşturan ürünleri satın alırım.		0.662				0.588
6. Alışveriş yaparken çevreye duyarlı (doğa dostu, geri dönüştürülebilir, hayvan haklarına saygılı vb.) ürünler satın alırım.		0.592				0.537
21. Yaşadığım şehirdeki geri dönüşüm tesisine atıklarımı gönderebilirim.		0.489				0.558
22. Geri dönüşüm sürecinde çok fazla enerji harcadığı için ambalaj atığı olmayan ürünleri tercih ederim.		0.383				0.607
11. Geri dönüşüm ürünlerini satın alırım.		0.358				0.725
15. Eskiye/yıpranan ürünleri farklı amaçlarla kullanabilirim.			0.796			0.307
18. Yıpranan ürünleri tekrar kullanmanın yolunu bulabilirim.			0.747			0.385
5. Kırılan veya bozulan ürünleri tamir edip tekrar kullanırım.			0.730			0.452
10. Eskimiş yıpranmış kıyafetlerimi onarıp tekrar kullanırım.			0.475			0.671
13. Atıklarım doğaya olan etkisini üzerinde düşünürüm.				0.691		0.424
17. Gündelik tüketim alışkanlıklarımızı gözden geçiririm.				0.542		0.601
3. Tüketim tercihlerimin etkisinin farkında olarak davranırım.				0.534		0.659
8. Tükettiğimiz her şeyin üretim yolculuğunu hakkında bilgi sahibiyim.				0.384		0.638
9. Takas yaparak ihtiyaçlarımı karşılarım.					0.783	0.354
23. Takas tekniğiyle ihtiyaçlarımı gideririm.					0.730	0.438
14. İkinci el pazarlarından alışveriş yaparım.					0.486	0.641
4. İkinci el eşya satın alırım.					0.438	0.738

Note. 'Principal axis factoring' extraction method was used in combination with a 'oblimin' rotation.

## Use of ASSURE MODEL in ELT: Reflections on the learning and teaching process

Hatun Vera Akşab<sup>1</sup>, Melike Özyurt<sup>2\*</sup>

<sup>1</sup>Gaziantep University, Institute of Educational Sciences, Gaziantep, Türkiye

<sup>2</sup>Gaziantep University, Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

### ARTICLE HISTORY

Received: Oct. 21, 2024

Accepted: May 31, 2025

### Keywords:

ASSURE instructional design model,

English language teaching,

Instructional technology,

Web 2.0 tools.

**Abstract:** Since 2020, the effects of the pandemic and advancements in instructional technologies have played a significant and transformative role in the educational system. The integration of student-centered activities has been emphasized by new technology, which continuously enhances educational processes. Additionally, the use of instructional design models, such as ASSURE, has been highlighted. This study aims to examine the reflection on the implementation of the ASSURE instructional design model in an English course within the learning-teaching process. In this case study, the 6th-grade English course was designed using the steps of the ASSURE instructional design model and implemented with 22 sixth-grade students during the spring term of the 2023-2024 academic year in Gaziantep, Türkiye. Data were collected through interviews with students and the researcher's diary, which were analyzed through content analysis. The findings indicate that the use of the ASSURE instructional design model in the English course enhanced students' academic achievement, increased their engagement, and positively influenced their attitudes toward the English language course.

## 1. INTRODUCTION

Taking the curriculum as a reference in the learning and teaching process is a fundamental determinant of educational efficacy. The curriculum serves as a critical framework that facilitates the attainment of educational objectives while simultaneously adapting to institutional requirements to ensure sustainability (Hadi, 2022). Foreign language instruction is a key area of the curriculum. The MoNE English Curriculum (2018) for 6th grade aims to foster creativity, imagination, and self-expression through the functional use of language in daily contexts. This approach encourages students to conceptualize language as a communicative instrument by utilizing authentic textual materials, including poetry, visual media, and conversational discourse, while fostering experiential learning through interactive pedagogical strategies such as gamification, dramatic performance, and puppet-mediated instruction. An eclectic approach that integrates diverse pedagogical methodologies, tailored to sociocultural contexts, is recommended for effective foreign language instruction (Tosun, 2012). The eclectic

\*CONTACT: Melike ÖZYURT ✉ [melike.ozyurt@yahoo.com](mailto:melike.ozyurt@yahoo.com) 📍 Gaziantep University, Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>



approach necessitates the integration of both visual and linguistic components of instructional materials deemed pedagogically appropriate. This enables educators to transcend reliance on purely verbal resources and instead leverage multimodal instructional strategies (Mwanza, 2017). Discussions on language pedagogy frequently engage with debates concerning the role of grammar in curriculum design, the selection of curricular frameworks, the teaching of communicative competencies, learner motivation, effective instructional strategies, vocabulary acquisition, and the integration of technological advancements in language instruction (Rodríguez-Izquierdo, 2021). The use of educational technology is instrumental in fostering learner engagement and optimizing language acquisition outcomes. To maximize efficiency, the selection of technological tools should consider factors such as portability, usability, and cost (Nagy, 2021). Moreover, educators must maintain awareness of emerging technological innovations and their pedagogical applications (Shadiev & Wang, 2022). Recent years have witnessed a paradigm shift in foreign language instruction, characterized by the adoption of novel educational models such as ‘computer-assisted learning’, ‘distance learning’, ‘flipped learning’, ‘adaptive learning’, ‘deep learning’, and ‘blended learning’ (Boyadzhieva, 2014; İrmiş & Uludağ, 2023; Jones, 2019; Lei, 2023).

Due to the diversity of instructional learners' needs, characteristics, and instructional goals, no single approach can fully meet all pedagogical needs. Accordingly, it is essential to identify an appropriate instructional model and design the implementation process based on pedagogical needs. Instructional design is defined in scholarly literature as the systematic application of pedagogical principles and learning theories to enhance instructional quality (Brown & Green, 2006). In other words, instructional design encompasses a systematic and reflective process aimed at fostering targeted knowledge and skill acquisition within specific learner demographics. The process entails transforming broad educational theories into practical teaching materials and methods (Gustafson & Branch, 1997). Instructional design can be realized through the application of various models. These models include ADDIE, ASSURE, Dick Carry and Carry, Kemp Morrison and Ross, Seels and Glasgow. The primary objective of instructional design models is to guide educators in structuring instruction rather than merely resolving isolated pedagogical issues (Molenda *et al.*, 1996). Consequently, instructional design models play a pivotal role in enhancing learning effectiveness and mitigating instructional challenges. These models offer structured frameworks for the systematic implementation of instructional strategies (Şimşek, 2017). The implementation of these models also enables the optimization of learning outcomes and performance in various educational contexts by formulating structured pedagogical plans (Smith & Ragan, 1999; Reigeluth, 1983; Reiser & Dempsey, 2008). The meticulous planning of instructional activities, coupled with the effective integration of technological tools such as digital screens, projectors, and online learning platforms, significantly influences instructional efficacy (Cooke, 2008). The integration of technological resources, including smartphones, personal computers, tablets, laptops, and smart boards, enables seamless access to information while simultaneously fostering learner motivation and collaborative engagement (Ghavifekr & Rosdy, 2015). In this context, the incorporation of instructional technologies into the teaching-learning process emerges as a critical consideration. One of the most widely recognized instructional design models that integrates technology is the ASSURE model.

### 1.1. ASSURE Instructional Design Model

The ASSURE model distinguishes itself among various instructional design frameworks utilized in educational practice through its distinctive focus on the systematic integration of technology in teaching and learning environments (Heinich *et al.*, 1999). The model provides a structured framework for incorporating technological resources into instructional planning to enhance learning experiences (Shelly *et al.*, 2012). The ASSURE model comprises six systematic stages: 1) analyzing learners, 2) stating objectives, 3) selecting media and materials,

4) utilizing media and materials, 5) requiring learner participation, and 6) evaluating and revising instructional strategies (Smaldino *et al.*, 2005).

The ASSURE instructional design model consists of six interrelated steps that collectively aim to enhance the instructional process. The first step, analyzing learners, underscores the need to evaluate individual input qualities, learning methodologies, and overall attributes (Heinich *et al.*, 1999). In other words, learner analysis, the first phase of the ASSURE model, focuses on identifying key learner attributes such as general traits, background knowledge, and learning preferences (Megaw, 2006). This is followed by stating objectives, which refer to goals students are expected to achieve by the end of the instructional process. In this step, both the desired behaviors and the conditions necessary for achieving these results are discussed (Kim & Downey, 2016). The creation of objectives focuses on educational goals rather than the methods of instruction; however, well-defined objectives assist practitioners in selecting media and resources and in the evaluation phase (Smaldino *et al.*, 2015). In this next step, selecting media and materials involves determining the most suitable approach, environment, and materials to meet the previously established objectives (Megaw, 2006). After assessing the learners and identifying instructional goals, the next step is to create a connection between these goals and the objectives. Consequently, the teacher must decide on the approach that best suits the learners and their learning objectives (Smaldino *et al.*, 2008).

Once the media and materials have been selected, the next phase—*utilizing media and materials*—is implemented. At this stage, the chosen or developed instructional resources are actively employed in the learning process. It is widely recognized that students derive the greatest benefit from materials that align with their individual learning styles, thereby enhancing their ability to achieve the intended learning outcomes. To maximize the effectiveness of media and materials, a range of instructional strategies should be integrated, including the incorporation of diverse technological tools (Heinich *et al.*, 2001). The following step emphasizes learner participation, recognizing that active student engagement in the learning process is essential for effective pedagogy and the attainment of established learning goals. This stage also suggests that educators should strive to keep learners engaged during instruction, allowing them to benefit from learning opportunities in an educational setting (Heinich *et al.*, 1999; Megaw, 2006). Finally, the process concludes with the *evaluation and revision* phase. In this final step, the instructional design is systematically reviewed, and necessary modifications are made. This includes assessing the extent to which students have achieved the learning objectives, analyzing levels of learner engagement, evaluating the effectiveness of the instructional materials, and identifying and addressing any gaps or areas for improvement (Heinich *et al.*, 2001).

The ASSURE instructional design model is widely utilized by educators in the development of learning activities and lesson planning (Russell *et al.*, 1994; Russell & Butcher, 1999; Smaldino *et al.*, 2015). Several factors contribute to its popularity. First, the model emphasizes the use of technology in instructional activities, accommodates short teaching durations, and supports individualized learning (Baran, 2010; Gündüzalp & Yıldız, 2020). Second, it effectively facilitates the integration of technology into educational settings (Kim & Downey, 2016; Shelly *et al.*, 2012). In recent years, external factors such as the COVID-19 pandemic and natural disasters (e.g., earthquakes) have further accelerated the adoption of technology in education, compelling teachers to adapt quickly to online teaching environments (Sun *et al.*, 2020). During the pandemic, two significant trends emerged: the rapid expansion of distance education and the acceleration of innovations in educational technologies (Ashour, 2021; Kang, 2021; Whitelock, 2024). Consequently, the ASSURE model has garnered increased attention as a relevant and adaptable framework for technology-enhanced instructional design.

## 1.2. Literature Review

The literature review reveals that extensive research has examined the impact of the ASSURE instructional design model on a range of educational outcomes, including the development of higher-order thinking skills, positive attitudes toward technology integration, environmental awareness, and the enhancement of social-emotional competencies. Additionally, the literature includes research studies exploring the implementation of the ASSURE model across a range of disciplines, such as mathematics, ICT, environmental education, music, science, and foreign languages.

Within this context, numerous studies have also investigated the effectiveness of the ASSURE instructional design model in enhancing students' higher-order thinking skills—particularly critical thinking and mathematical communication—within the field of mathematics education. Findings from these studies indicate that the ASSURE model is effective in fostering critical thinking skills among secondary school students (Kristianti *et al.*, 2017) and significantly contributes to the development of mathematical communication skills (Sundayana *et al.*, 2017).

Researchers have also explored the contribution of the ASSURE model to technology-integrated instruction, particularly its role in shaping students' attitudes toward information and communication technologies (ICT), reducing classroom anxiety, and addressing individual learning differences. For instance, Gündüzalp and Yıldız (2020) investigated the effects of an ICT course designed using the ASSURE model on students' attitudes toward ICT and their perceptions of the course. Çibir and Yazgan (2021) developed a lesson on 'Addition by Mind' for second graders using the ASSURE model. Their findings revealed that courses structured with the ASSURE instructional design model fostered positive attitudes toward computer courses while reducing student anxiety (Gündüzalp & Yıldız, 2020). Additionally, this model improved the effectiveness of teaching by mitigating learning challenges associated with individual differences (Çibir & Yazgan, 2021).

The ASSURE instructional design model has been applied across a range of disciplines, including environmental education, music, and science, with studies highlighting its effectiveness in promoting academic achievement, fostering social skills, and enhancing instructional efficiency. Çatar and Özdilek (2023) examined the effects of implementing the ASSURE instructional design model in environmental education on middle school students' environmental attitudes. Their findings indicated a significant improvement in students' attitudes toward the environment as a result of instruction based on the ASSURE model. İrmış & Uludağ (2023) examined the effects of learning environments developed by integrating the ASSURE instructional design model, the station learning method, blended learning, mobile games, and Web 2.0 tools on both students and teachers in the context of teaching basic music theory. Their findings indicated that the implementation of the ASSURE model enhanced students' independent learning, motivation, cooperation, communication, interaction, socialization, and productivity skills. Notably, research on lesson planning grounded in instructional design models is predominantly concentrated in the context of science education. These studies have demonstrated that the ASSURE model significantly enhances students' academic achievement in science courses (Kaya *et al.*, 2020) and provides various benefits for teachers, including time efficiency, improved student performance, increased student engagement, and the development of students' computer skills (Karadeniz & Karamustafaoğlu, 2022).

Within the context of English language instruction, research specifically addressing the ASSURE instructional design model remains limited. Altın (2021) evaluated the effectiveness of English language teaching practices informed by the ASSURE model and found that it significantly enhanced student achievement and engagement, while also emphasizing the need for ongoing evaluation and refinement of instructional practices. Similarly, Zai *et al.* (2024) explored the model's impact on student motivation, concluding that its structured and

interactive design fosters student engagement and contributes to more effective English language learning experiences. Adedapo and Opoola (2021) highlighted that the ASSURE model improves the quality of English language instruction by helping teachers align lesson objectives with learner needs and by encouraging the effective use of instructional media, which fosters a supportive and engaging learning environment. Sezer *et al.* (2013) emphasized that the learner-centered ASSURE model, grounded in cognitive learning theories, enhances English language teaching by integrating technology in a structured way that supports diverse learning styles and encourages active student participation. Kazancı *et al.* (2020) reported that the ASSURE instructional design model enhances English language teaching by promoting technological literacy, fostering active student participation, and supporting the development of customized lesson plans tailored to learners' needs.

The shift to online learning prompted by the COVID-19 pandemic has profoundly impacted educational practices, as reflected in the growing body of research on instructional design models since 2019 (Hu & Huang, 2022; Maican & Cocoradă, 2021). Technological advancements have proven particularly beneficial in foreign language education, especially in English language teaching, by improving access to resources, enabling personalized and interactive learning experiences, and increasing student engagement (Mulya & Putro Setyo, 2024; Poloju, 2024). In light of the increasing demand for technology-enhanced instruction, the deliberate integration of digital tools into lesson planning has become essential (Asandaş & Hacicaferoğlu, 2021; Çetinkaya & Taş, 2016; Shelly *et al.*, 2012; Sezer *et al.*, 2013). As technology continues to expand the scope of educational possibilities, educators must develop a strong understanding of instructional design that effectively incorporates digital tools (Marín *et al.*, 2018).

Against this backdrop, the present study explores the reflections of both students and the researcher on the implementation of the ASSURE instructional design model in the 6th-grade English unit 'Saving the Planet.' Addressing a gap in the literature, this study contributes to the field by providing applied insights into the use of the ASSURE model within a specific unit of English language teaching—an area that remains underexplored. Its significance lies in offering an instructional design framework grounded in the ASSURE model and evaluating its effectiveness through student feedback.

## 2. METHOD

### 2.1. Research Design

This study was designed as a qualitative case study, one of the widely employed research approaches in educational sciences. Case studies are utilized to explore contemporary phenomena within their real-world context (Stake, 1995). This approach is especially valuable when the boundaries between the phenomenon under investigation and its context are indistinct, and when a comprehensive understanding necessitates the use of multiple data sources (Yin, 1984). In case study research, the researcher examines one or more bounded systems in depth, utilizing multiple data collection instruments, such as observations, documents, interviews, audiovisual materials, and reports (Creswell, 2007). Various case study designs exist, each serving distinct analytical purposes. A single-case study focuses on one unit of analysis, such as an individual, program, institution, or school (Yıldırım & Şimşek, 2008). Within this framework, the holistic single-case design is particularly suitable for in-depth exploration of a specific case in its entirety (Storey, 2007). Given the scope and context of the present study—examining a 6th-grade English language course structured using the ASSURE instructional model—the holistic single-case design was identified as the most appropriate methodological approach.



## 2.2. Participants

The study group comprises one English language teacher and 22 sixth-grade students from a state-affiliated secondary school in Gaziantep/ Türkiye, during the spring term of the 2023-2024 academic year. Of the 22 students, 13 were female and 9 were male. The participants were selected through purposive sampling based on the school's readiness for technology-integrated instruction and voluntary participation. The students came from diverse socio-economic backgrounds and had limited prior experience with Web 2.0 tools.

The English language teacher, who is one of the researchers, has 12 years of teaching experience and holds a master's degree in Curriculum and Instruction. She was responsible for lesson implementation and reflective documentation through a researcher diary.

## 2.3. Data Collection Tools

The data collection tools employed in this study included a student interview form and the researcher's diary. To enhance the validity of the research in case studies, utilizing multiple data sets is considered crucial (Yin, 2003). In the initial development of the interview form, the researchers crafted a draft consisting of 13 questions. Subsequently, the draft was reviewed by two experts, whose feedback led to the elimination of certain questions (e.g., "Do you have any suggestions for improving this implementation?", "Does this implementation affect classroom interaction", and "Would you like see similar implementations in other units?") and the modification of the some statements (such as replacing "use of educational technologies" with "use of technology" and "technology-based activities and applications" with "ASSURE model."). Following this, the input of a language expert was sought, and the form was prepared for a pilot implementation. However, based on feedback from interviews with several students during the pilot study, it was determined that some questions should be discarded due to a lack of clarity; students were unable to provide meaningful responses to these queries. Consequently, the final version of the student interview form was refined to consist of 10 questions. These questions aimed to explore students' perceptions regarding the distinctions between ASSURE-supported courses and traditional implementations, the contributions and limitations of the ASSURE model to the learning process, and its potential applicability across other disciplines.

The researcher's diary served as the second data collection tool. The researcher meticulously documented the events during the study, incorporating various data sources, such as observations, direct quotations, brief notes, and personal reflections. Researcher diaries play a pivotal role in qualitative research, enriching both the research process and its outcomes, thereby enhancing overall effectiveness and depth (Borg, 2001; Mills, 2003; Gerg, 2009). In this study, the researcher's diary was utilized to record detailed observations of the classroom environment throughout the implementation of the ASSURE model.

## 2.4. Design and Implementation Process of the ASSURE Instructional Model

This section of the study addresses the implementation and processing of the topic "Saving the Planet", which is the 9<sup>th</sup> unit of the 6th-grade English course, following the steps outlined in the ASSURE instructional design model. Before the main implementation, the ASSURE model was piloted in another classroom with similar characteristics. Feedback and observations from the pilot implementation were taken into account to refine the process for the subsequent classroom implementation. For instance, some Web 2.0 tools that served similar purposes were removed from the list after being identified during the pilot study. Following the pilot phase, the instructional design was implemented with the study group over a three-week period. A detailed account of the instructional design development process, structured according to the steps of the ASSURE model, is provided below.



### 2.4.1. Analyze learners

This phase, which constitutes the first stage of the ASSURE model, involves an analysis of the students' general characteristics, prior knowledge, and learning styles.

*General Characteristics:* The study group consisted of 22 sixth-grade students (9 males and 13 females) aged 11-12 years. All students demonstrated fundamental academic competencies, including hearing, seeing, writing, and reading. No students required individualized education, as there were no identified learning disabilities or special educational needs. In terms of socioeconomic background, the majority of students belonged to middle-income families. Additionally, all students had access to desktop computers, laptops, or tablets, enabling them to engage with technology-based learning applications beyond the classroom environment.

*Prior Knowledge:* According to the English curriculum, a learning outcome from the “Saving the Planet” unit in sixth grade has been incorporated into the “Health” unit in the fifth grade. However, an assessment of students' prior knowledge reveals that their level of readiness for this unit is insufficient.

*Learning Styles:* Based on the learning styles inventory administered by the school guidance service at the beginning of the academic year, it was determined that students predominantly exhibit auditory, visual, and kinesthetic learning preferences. Consequently, the instructional design was planned to incorporate diverse learning activities that cater to all these learning styles, ensuring an inclusive and effective learning experience.

### 2.4.2. State objectives

At this stage, the objectives outlined in the curriculum are prioritized. In this study, the unit objectives were determined based on the four language skills specified in the 2018 English curriculum.

### 2.4.3. Select methods, media and materials

As emphasized in the relevant curriculum, an action-oriented teaching approach was adopted in this study. This approach underscores the importance of student-centered learning, promoting active participation and opportunities for students to express themselves in the target language. Additionally, it highlights the integration of technology and individualized learning in digital environments, leveraging the advantages offered by multimedia tools and materials. The selection of methods, approaches, and resources was guided by their cost-effectiveness, pedagogical benefits, and ease of use for teachers. All activities were conducted within the classroom setting due to the availability of necessary resources. The methods, techniques, and materials utilized in the study are outlined below:

*Smart Board and Education Informatics Network (EBA) Applications:* The smart board was actively used throughout the instructional process. The EBA application and its digital content were integrated to facilitate the attainment of unit objectives. Additionally, teachers could adapt and utilize EBA resources for their instructional needs. EBA provides a wide range of educational content, including interactive activities and assessments. In this study, activity-based questions available in EBA were employed for evaluation purposes.

*Google Classroom:* Google Classroom application was utilized to create lessons, deliver content, and enhance vocabulary instruction aligned with the unit objectives. This platform also facilitated teacher-student communication, assignment tracking, and peer interaction. Students could submit assignments, view their peers' work, and provide comments, while teachers offered feedback and shared written materials related to the topic.

*Web 2.0 Tools:* A variety of Web 2.0 tools were integrated into the instructional process to enhance student engagement and facilitate both in-class and out-of-school learning. The selected tools included “Word Art”, “Cram”, “Bamboozle”, “Word Cloud”, “Mentimeter”, “Learning Apps,” and “Canva.” These resources provided interactive and creative learning

opportunities tailored to different learning styles. Canva enables students to design visual material such as posters, banners, slides, and infographics. With Learning Apps, students reinforced the concepts covered in the lesson through a variety of interactive exercises and educational games. Students are able to play instructional games with the Bamboozle tool. The Word Art, Word Cloud, and Mentimeter allowed students to create visually appealing word clouds, helping them organize and internalize vocabulary in a dynamic way.

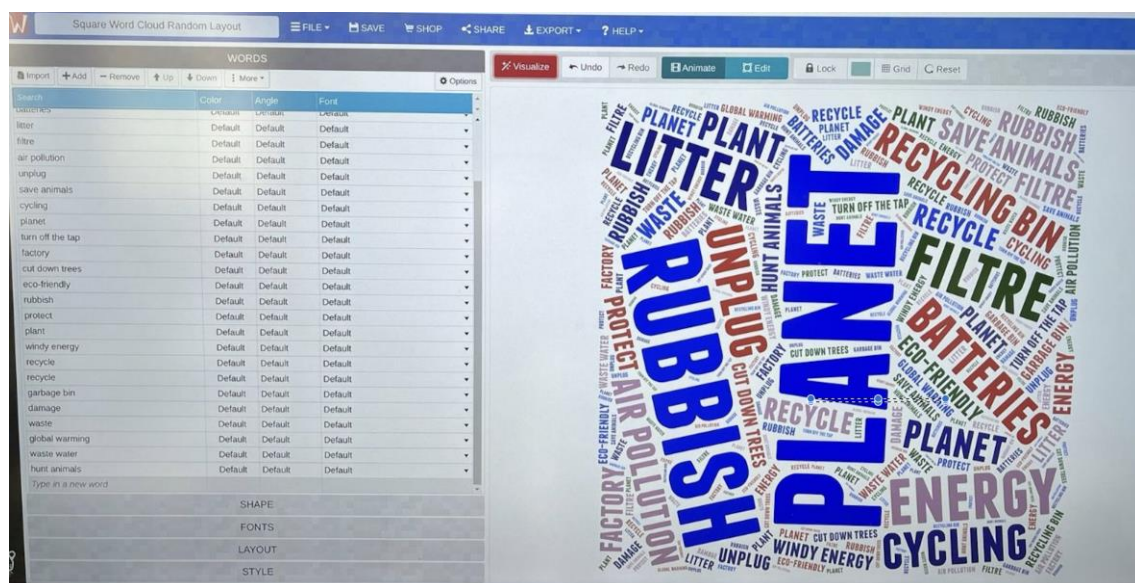
*Textbook and Other Supplementary Resources:* MoNE Textbook and other additional supplementary materials (such as quizzes as written assessment materials, as well as unit achievement tests and skills-based assessments prepared by the MoNE) were utilized throughout the unit to provide structured content and support student learning.

*Song:* As part of the instructional design, the song “Protect Our Planet” was incorporated to align with the unit objectives, reinforcing key concepts through music-based learning.

*Quiz as Written Material:* In addition to the activity-based questions available in the EBA application, the quizzes developed by the MoNE, comprising achievement comprehension tests and skill-based questions, were employed as written assessment tools to evaluate student progress.

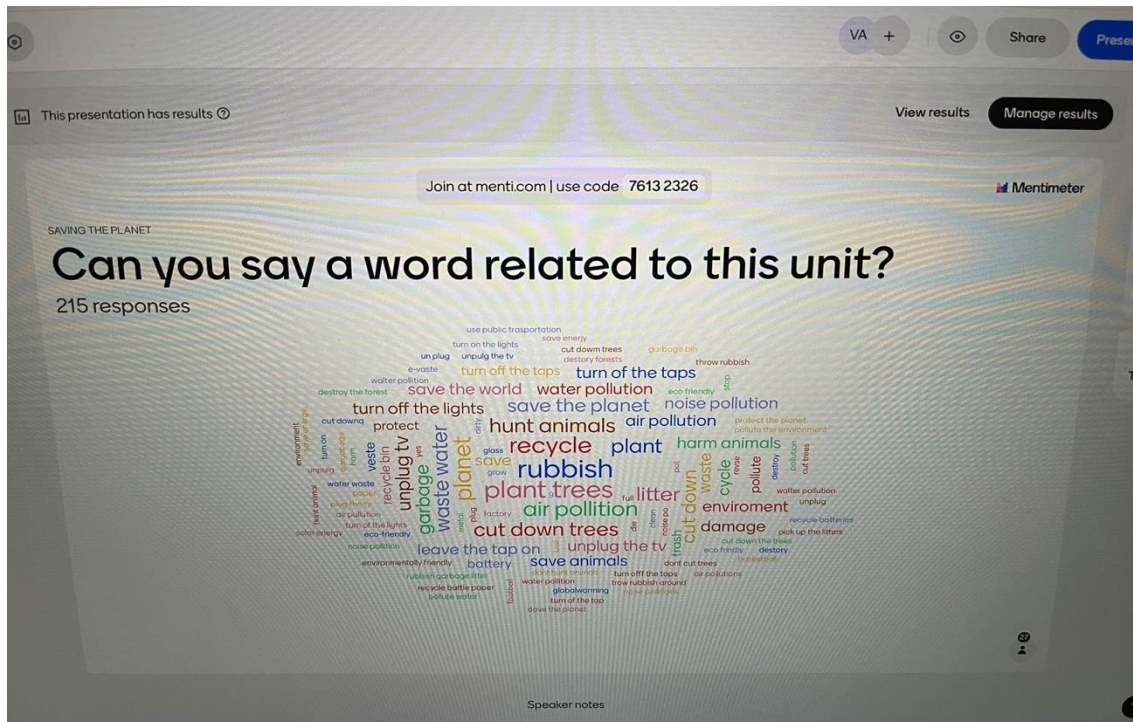
#### 2.4.4. Utilize media materials and require learner participation

*First Week:* Before the lesson, a vocabulary list containing the English and Turkish meanings of the key terms from the unit was shared with the students via Google Classroom. Students were instructed to review the vocabulary before attending the class to ensure a foundational understanding. In-class activities were designed to reinforce vocabulary acquisition among students who had familiarized themselves with the words in advance. Initially, flashcards created using the Cram tool were presented to students, who were expected to observe each card, infer its meaning, and articulate the corresponding word. To further reinforce vocabulary learning, an interactive game featuring the newly introduced words was conducted. This activity offered students an additional opportunity to engage with the target vocabulary in an interactive and enjoyable manner. Subsequently, Word Art and Word Cloud tools were utilized to create a visual representation of unit-related vocabulary. Each student contributed by recalling and stating a word associated with the unit, culminating in the collaborative formation of a word cloud.



**Figure 1.** *Creating a word cloud with the Word Art tool.*

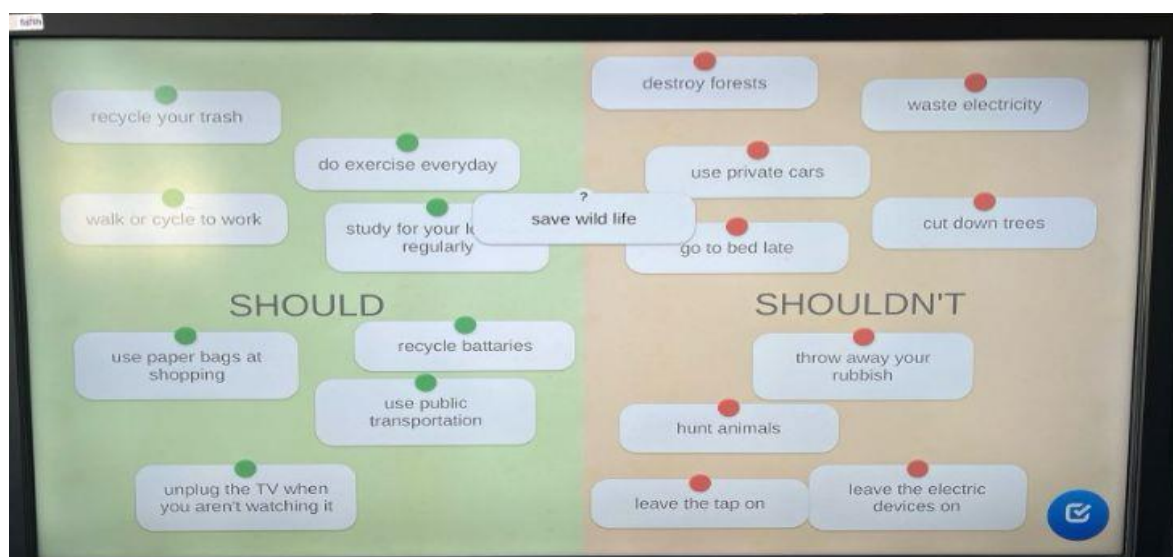
Additionally, students reinforced their vocabulary acquisition using the Bamboozle tool. Through this platform, they engaged in various interactive activities, such as matching English words with their Turkish equivalents, identifying word meanings in a wheel of fortune game, and participating in other vocabulary-based exercises offered by the tool. At the end of the lesson, students were assigned a Mentimeter activity as homework. The Mentimeter code was shared via Google Classroom, and students were instructed to enter newly learned words into the tool. This activity encouraged active recall and consolidation of vocabulary. To further reinforce their learning, students watched EBA's instructional videos related to the unit vocabulary, allowing them to review and solidify their understanding through multimedia content.



**Figure 2.** *Creating a word cloud with the Mentimeter tool.*

*Second Week:* Before the lesson, written material on the ‘should/shouldn’t’ grammar structure was shared with students via Google Classroom. Students were expected to review the material in advance to ensure preparedness for the lesson. During the lesson, the topic was first introduced and explained, followed by reinforcement activities using EBA content. Students watched instructional videos related to the topic and participated in accompanying exercises to deepen their understanding. Given the significance of game-based learning in foreign language instruction, Web 2.0 tools that integrate technology-enhanced games were incorporated into the lesson. Learning Apps was the first tool utilized, offering a range of interactive activities related to the should/ shouldn't structure. To further consolidate learning, a quiz was conducted at the end of the session, allowing students to internalize the concepts through engaging and structured practice.





**Figure 3.** “Should” activity with the Learning Apps tool.

This part of the lesson was devoted to developing speaking skills. To initiate the activity, visuals related to the predetermined environmental theme were placed in envelopes. Each student selected an envelope and was instructed to construct a sentence using the structure ‘We should.....’ based on the visual they received. For homework, students were assigned a poster creation activity using the Canva tool. They designed environment-themed posters to reinforce their understanding of the topic creatively. Upon completion, students uploaded their posters to Google Classroom, where they could view and engage with their classmates’ work, fostering peer interaction and collaborative learning.



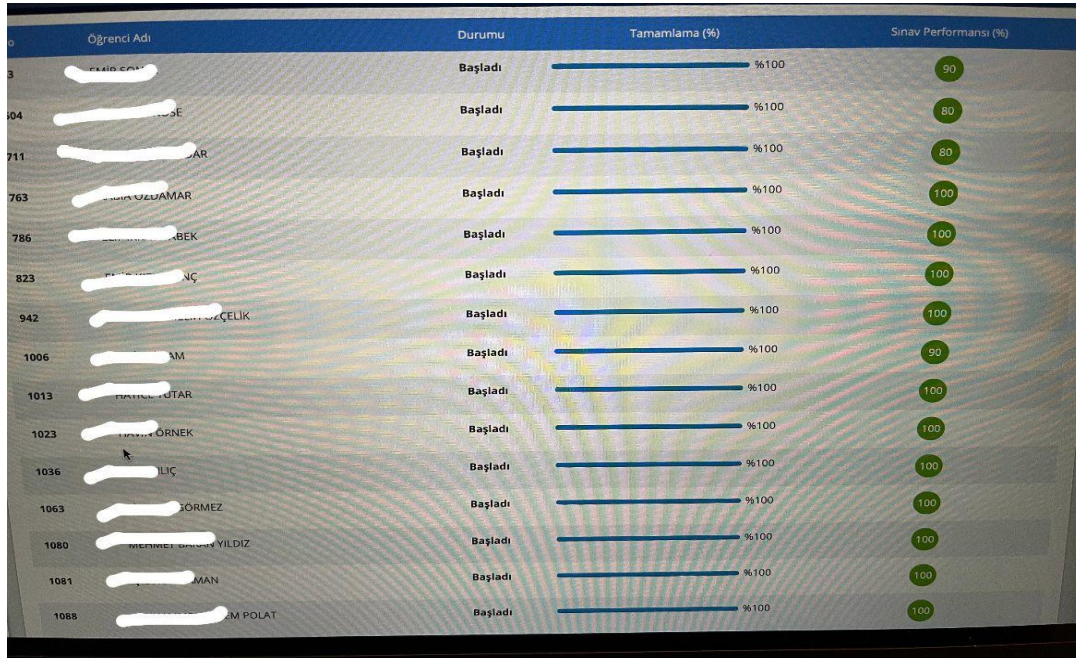
**Figure 4.** Making a poster with the Canva tool.

*Third Week:* As part of the unit, the song ‘Saving the Planet’- which was easy for students to follow and sing along with- was introduced. Initially, students listened to the song multiple times to familiarize themselves with its lyrics and rhythm. Subsequently, they both listened to and sang the song, enabling them to naturally acquire and internalize key vocabulary and expressions related to the unit. To assess students’ understanding and retention of the unit content, a written exam was administered at the end of the unit. Additionally, tests and exercises from EBA were utilized to evaluate whether the topics had been effectively reinforced. Finally, students completed a MoNE assessment, which included unit achievement tests and skill-based questions, providing a comprehensive measure of their learning outcomes.

#### 2.4.5. Evaluate and revise

In the study, various assessment tools were utilized, including ready-made questions and exercises from EBA, quizzes as written assessment materials, as well as unit achievement tests and skills-based assessments prepared by the MoNE. The integration of EBA’s evaluation

resources not only streamlined the assessment process but also enhanced practicality for teachers, as EBA generates an automated analysis report upon completion of the exams. Moreover, EBA's detailed evaluation reports provided valuable insights into student performance, enabling a more comprehensive assessment of learning outcomes.



O	Öğrenci Adı	Durumu	Tamamlama (%)	Sınav Performansı (%)
3	FAHİR ÇOK	Başladı	%100	90
104	...	Başladı	%100	80
711	...	Başladı	%100	80
763	...	Başladı	%100	100
786	...	Başladı	%100	100
823	...	Başladı	%100	100
942	...	Başladı	%100	100
1006	...	Başladı	%100	90
1013	...	Başladı	%100	100
1023	...	Başladı	%100	100
1036	...	Başladı	%100	100
1063	...	Başladı	%100	100
1080	...	Başladı	%100	100
1081	...	Başladı	%100	100
1088	...	Başladı	%100	100

**Figure 5.** Evaluation of the activity sent from EBA.

The evaluation of the instructional process indicated that the implementation of the instructional design was carried out smoothly, with students demonstrating active engagement in the learning activities and the instructional materials effectively fulfilling their intended objectives.

## 2.5. Data Collection Process

In the context of the research, semi-structured interviews were conducted with the students to gather their perspectives on the instructional design developed based on the ASSURE instructional design model. Following the implementation of the instructional design, interview questions were asked at the end of the research process to evaluate the effectiveness of the approach from the students' viewpoints. In order to enable students to express themselves more effectively, the interview form was prepared in Turkish. Rather than interviewing the entire class, a sample of nine students was selected, with three students chosen from each learning style group. Each interview session lasted between 20 to 25 minutes, and the questions from the interview form were presented sequentially. The interviews were conducted neutrally, ensuring that the researchers did not influence the participants' responses.

In addition to the interviews, a researcher's diary was employed as another data collection tool. Throughout the implementation process, the first researcher maintained a reflective diary to document observations and key developments. During this phase, the ASSURE instructional design practices were analyzed with respect to their impact on the teaching and learning process, the learning environment, and the integration of instructional technology. The determination of these dimensions was guided by a comprehensive approach, ensuring that the analysis complemented, rather than duplicated, the findings from previous literature on ASSURE instructional design practices. Additionally, expert opinions and the questions used in the student interviews played a decisive role in shaping the evaluation framework.



## 2.6. Data Analysis

In this study, the data obtained from semi-structured interviews and the researcher's diary were analyzed using content analysis. The analysis process followed a systematic sequence, beginning with coding the data, followed by identifying themes, structuring and defining the data based on these codes, and finally interpreting the findings. To ensure rigor and reliability, the data were read twice, with significant observations noted, after which the researchers independently coded the data without prior collaboration. Following this, similar codes were grouped, and overarching themes were generated. Unlike descriptive analysis, no predefined conceptual framework guided the process at the outset. Instead, in line with content analysis methodology, an inductive approach was adopted, allowing codes and themes to emerge naturally from the data rather than being predetermined (Strauss & Corbin, 1990). For confidentiality, participants were anonymized during the interview analysis; they were assigned coded identifiers (e.g., Student 1 (S1), Student 2 (S2)). To enhance the clarity and comprehensibility of the findings, the identified themes and codes were visualized using MindMeister, with corresponding figures created to systematically represent the results.

## 2.7. Reliability and Validity

To ensure the validity and reliability of the research, several measures were implemented. Firstly, to enhance the internal reliability, all findings were presented directly and without interpretation, ensuring transparency in data reporting. Participants' responses were included as direct quotations to provide authentic insights into their perspectives. Efforts were made to maintain objectivity in both the interpretation and description of the data. Additionally, the data coding process was conducted independently by the researchers to avoid bias. To establish the reliability of the research, the researchers later convened to discuss and reconcile the identified codes and themes. Miles & Huberman's (1994) reliability formula ( $\text{Reliability} = \frac{\text{Agreement}}{\text{Agreement} + \text{Disagreement}}$ ) was applied, yielding an agreement rate of .87, which indicates a high level of reliability. Following this calculation, the researchers further deliberated on discrepant codes, achieving full consensus (100%) through discussion. To ensure internal validity, the researchers actively participated in all stages of the research, including data collection and analysis. The first researcher played a key role in implementing the instructional design and maintaining the researcher's diary. To enhance external validity, a detailed account of the research process was provided, outlining each step undertaken during the study. Finally, semi-structured interviews and the researcher's diary were employed to enhance the validity, ensuring the use of multiple data sources. By incorporating these diverse sources, data triangulation was achieved, allowing for a more comprehensive and holistic understanding of the research findings.

## 3. RESULTS

### 3.1. Student Interviews

As a result of the interviews with the students, five themes and codes related to each theme emerged. The themes and codes are presented in Figure 6. As seen in Figure 6, the first theme is centered around what the ASSURE instructional design model evokes in students. The related codes include "various activities", "use of technology (Web 2.0 tools)" such as "Canva", "Learning apps", "Word Cloud", "Cram", "Mentimeter", "Google classroom", and use of "EBA." One student (S6) expressed that *"When I think of ASSURE model, I think of the games we played in Learning apps, the posters we created in Canva, and many different activities we did in the classroom"*. Another student (S3) mentioned that *"The activities we did in class and the song we sang are the first things that come to my mind."* The second theme highlights "ASSURE's contribution to the learning process", with codes such as "learning vocabulary easier, feedback, use of technology, permanent learning, being active, learning with fun, self-confidence, application of learning, increasing willingness/motivation, learning better, learning

faster, learning with activities and learning easier". One of the students (S1) stated that *"Before, I had difficulty understanding the topics, but now I understood the lesson better and realized that I did not forget what I learned."* Another student (S9) remarked that *"Learning words has never been so easy. When I see it, I immediately remember what it means."*



**Figure 6.** Reflections of ASSURE to learning-teaching process.

The third theme addresses the differences of ASSURE-supported courses from a routine course, identifying codes like “using technology”, “learning writing less”, “different activities”, “entertaining”, “memorability”, and “increasing interest.” One student (S7) noted: *“Yes, it was different from the English lessons we used to teach before because in our other lessons we did more writing, there was a lot of lecturing. But now we used technology more and did different activities.”* The fourth theme focuses on the compatibility of ASSURE applications with other courses, with codes including “Turkish, Social Studies, Maths and Science”. One student (S8) commented: *“I think this model is more suitable for Turkish and Social Studies lessons because these lessons are based on rote learning and should be learned by having fun.”* S2 stated that *“More suitable for Maths and Science. These lessons are difficult to understand, and being active and doing activities like these can make it easier for us to understand.”*

Finally, the last theme examines the disadvantage of the ASSURE application, where “noise” is the sole code. This is the only drawback mentioned by students regarding this issue. One of the students (S5) pointed out, *“When the smart board is switched on, there is usually a lot of noise in the classroom.”* One of the participants (S4) stated that *“Normally the whole class is silent while writing, but in this model, when there was no writing and technology was at the forefront, the noise broke out.”*

### 3.2. Researcher Diary

In the researcher’s diary, salient and meaningful aspects of the study were documented over a three-week observation period. The themes that emerged from the reflective diary pertained to the implementation of the ASSURE instructional design model in English language teaching (ELT) and were categorized as “Web 2.0 tools,” “students,” and “classroom environment.” These themes, along with their associated codes, are presented in Figure 7.



**Figure 7.** *Researcher Diary.*

Concerning Web 2.0 tools, it was noted that students were initially unfamiliar with the concept; however, they engaged with and learned to utilize various tools throughout the process. During the implementation, students demonstrated recognition of tools such as Learning Apps, Canva, Mentimeter, Word Art, and Google Classroom. Notably, students expressed a strong preference for Canva, particularly enjoying the process of creating their own posters. Additionally, educational games facilitated through the Learning Apps tool were particularly popular among students. There was a marked increase in the active and conscious use of the EBA platform, contrasting with the previous behavior of students who only intermittently searched for homework assignments.

In terms of student engagement, a positive shift in attitudes toward the English course was observed overall. Even students who had previously shown little interest in the lesson became more active and engaged during classroom activities. Notably, students were able to understand the subject matter more meaningfully, and their self-confidence increased as they participated more actively in lessons. The integration of digital tools contributed to heightened interest in both the lesson and the subject matter, particularly facilitating easier retention of vocabulary. Furthermore, students expressed enjoyment in utilizing Web 2.0 tools, which facilitated a more enduring understanding of the material through diverse activities. They developed greater awareness of their language skills, particularly while creating posters with Canva and engaging in singing activities related to the unit. The students' natural curiosity about digital technology led them to find the lesson both enjoyable and motivating.

Lastly, regarding the classroom environment, the small class size was noted to enable a variety of activities. The classrooms were adequately equipped with technological resources, such as smart boards. It was observed that the implementation of the activities using the smart board resulted in a higher noise level compared to traditional lessons. Students tended to be quieter when writing on a blackboard; however, the increased use of digital tools in this model led to

more active and participatory classroom dynamics, which contributed to heightened enthusiasm and, at times, restlessness among students during lessons.

#### 4. DISCUSSION

This study explored the impact of the ASSURE instructional design model on the teaching and learning process, drawing on data from post-implementation student interviews and the researcher's reflective diary. The findings revealed consistent themes highlighting both the pedagogical and emotional effects of the model.

Firstly, the use of Web 2.0 tools and technology-integrated materials, such as Canva, Learning Apps, Mentimeter, Google Classroom, and EBA, emerged as central to students' engagement and learning. Students primarily associated the ASSURE model with the use of digital tools and engaging activities, indicating that technology integration not only supported vocabulary retention and feedback but also enhanced the accessibility and enjoyment of learning. The incorporation of tools such as Canva contributed to varied and visually appealing instruction, increasing student interest (Sari & Hasanah, 2022). These findings align with those of Smaldino *et al.* (2008) and Kim and Downey (2016), who emphasized that the ASSURE model improves instructional quality by promoting effective media integration and active student participation. Additionally, Juan (2011) emphasizes the model's systematic structure in applying instructional technologies, which was evident in the structured digital activity design implemented in this study.

The model's impact extended beyond cognitive gains to affective domains. Students reported increased motivation, self-confidence, and willingness to participate in class. Relevant studies show that the implementation of the ASSURE model has been shown to increase student engagement through the use of technology-based learning environments (Eliana *et al.*, 2024). The ASSURE model leads to better learning contexts (Zahran, 2023). These findings are in alignment with Arriyani & Pratama (2021) and Zhu *et al.* (2023), who found that ASSURE-based virtual instruction supports both engagement and language proficiency, while also providing valid structures for measuring student involvement. Likewise, Alhenaki & Alarfaj (2020) emphasize the role of the ASSURE model in enhancing learning motivation at intermediate levels. The ASSURE model positively impacts the learning process (Saputra *et al.*, 2021) by increasing students' engagement (Eliana *et al.*, 2024) and motivation (Kazanci *et al.*, 2020; Zai *et al.*, 2024).

The students clearly distinguished the ASSURE-supported lessons from routine ones. Unlike traditional methods focused on writing and lecturing, the ASSURE-based instruction provided multimodal, visually enriched, and activity-based experiences. This contrast aligns with Giang *et al.* (2022), who noted that the model integrates cognitive, emotional, and participatory dimensions of engagement both inside and outside the classroom. Moreover, students expressed a desire to see the ASSURE model applied in other subjects, especially Social Studies, Turkish, Maths, and Science. While Turkish and Social Studies were perceived as overly reliant on rote memorization, Mathematics and Science were viewed as more comprehensible when taught through activity-based methods. In a study by Şahbaz *et al.* (2024), students reported that they found educational games enjoyable and particularly well-suited to Science courses. The findings indicated that incorporating educational games not only enhanced learning but also made the process more engaging and enjoyable for students.

Math and Science were viewed as more accessible through activity-based approaches. This perspective finds support in Kristianti *et al.* (2017) and Sundayana *et al.* (2017), who demonstrated the model's capacity to enhance academic achievement in Math and Science. Similarly, studies have shown the effectiveness of ASSURE-based instruction in environmental education and science content (Çatar & Özdilek, 2023; Kaya *et al.*, 2020; Karadeniz & Karamustafaoğlu, 2022).

Finally, results from the researcher's diary provided observational evidence that reinforced student views. Students were initially unfamiliar with Web 2.0 tools, but through consistent use, they became more autonomous and enthusiastic participants. The model fostered an interactive classroom environment, where noise levels increased due to active learning rather than distraction. While this was seen as a limitation by some students, it was interpreted in the researcher's diary as a sign of heightened engagement. İrmış & Uludağ (2023) acknowledged that the dynamic nature of ASSURE-based environments contributes to student motivation and collaboration, even if it challenges conventional notions of classroom order.

## 5. CONCLUSION

This study demonstrated that the ASSURE instructional design model positively influenced English language learning by enhancing student engagement, motivation, and achievement. The integration of Web 2.0 tools and student-centered methods provided a dynamic learning environment tailored to various learning styles, resulting in more effective and enjoyable instructional experiences.

The findings underscore the model's potential to support both cognitive and affective outcomes in language education. Despite the study's contextual limitations, it offers practical insights for applying the ASSURE model in technology-integrated classrooms and suggests its broader applicability across disciplines, fostering active, reflective, and learner-focused instruction.

### 5.1. Pedagogical Implications

Based on the research findings, it is recommended that the ASSURE instructional design model be utilized in English language teaching and adapted for broader application across various subjects. Educators adopting the model should begin with a thorough analysis of students' age, interests, and learning preferences to design and deliver effective lessons. Successful implementation also requires the careful selection of appropriate instructional techniques, strategies, and resources, along with the establishment of clear and measurable learning objectives. Furthermore, designing engaging, practical activities that promote active student participation and effectively integrate technology is essential.

Teacher competency is crucial for the design and implementation of ASSURE instructional design. Therefore, to enhance teachers' pedagogical knowledge and skills, it is recommended that a comprehensive in-service training be provided, focusing on technology integration, learning analysis, and differentiated instructional strategies. Consequently, curricula and teacher training programs should be designed to support adaptive, innovative, and technologically advanced teaching methods.

### 5.2. Limitations

Although the study provides valuable insights into the application and pedagogical implications of the ASSURE instructional design model in technology-integrated English language teaching, it is not without limitations. Firstly, it exclusively focused on the implementation of ASSURE instructional design within a 6th-grade English classroom. Secondly, the study is constrained to the specific Web 2.0 tools employed during the implementation of the ASSURE model. Lastly, the study was conducted in a public-school setting, which imposes limitations related to the facilities and conditions typical of such environments.

The findings highlight the potential of the ASSURE model for fostering student-centered learning environments. To enhance the generalizability and adaptability of the model, it is essential to conduct comparative studies that assess its efficacy across diverse age groups and subject areas. Additionally, qualitative research is recommended to explore the challenges instructors face when implementing this model, along with their suggestions for addressing these challenges. Furthermore, a deeper understanding of the ASSURE model's role in digital learning environments could be achieved through long-term studies that investigate how



technology-supported instructional methods might be refined within the framework of the model. Finally, a comprehensive evaluation of the model should include an examination of its impact on affective outcomes, such as motivation, self-confidence, and attitudes, in addition to student achievement.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Gaziantep University Ethics Committee, 02.04.2024- 474996.

### Contribution of Authors

**Hatun Vera Akşab:** Literature review, Implementation, Visualization, Data Analysis and Reporting and Editing. **Melike Özyurt:** Methodology, Data Analysis, Validation, Reporting and Editing.

### Orcid

Hatun Vera Akşab  <https://orcid.org/0009-0003-7756-4752>

Melike Özyurt  <https://orcid.org/0000-0003-4527-9343>

### REFERENCES

- Adedapo, A., & Opoola, B.T. (2021). Levels of integrating the ASSURE model in lesson delivery of selected primary school teachers in Nigeria. *Journal of Language Teaching and Research*, 12(1), 177–182. <https://doi.org/10.17507/JLTR.1201.19>
- Alhenaki, M.S., & Alarfaj, A.M. (2020). The effect of teaching using ASSURE model on motivation toward learning in the intermediate school. *Journal of Educational and Psychological Sciences*, 4(2), 38-56. <https://doi.org/10.26389/AJSRP.M230719>
- Altın, M. (2021). Evaluation of the effectiveness of English language instruction based on the ASSURE model. *E-International Journal of Educational Research*, 12(5), 195-211. <https://doi.org/10.19160/e-ijer.1018149>
- Arriyani, N., & Pratama, P. (2021). English virtual based learning: Integrating technology and learning media through “ASSURE” teaching model. *Journal Pendidikan Bahasa Inggris*, 10(2), 421-429. <https://doi.org/10.26618/exposure.v10i2.6054>
- Asandaş, N., & Hacıcafareoğlu, S. (2021). Koronavirüs (Covid-19) döneminde uzaktan eğitim süreci [Distance education process in the Coronavirus (Covid-19) period]. *Mustafa Kemal Üniversitesi Eğitim Fakültesi Dergisi*, 5(7), 213-223.
- Ashour, S. (2021). How COVID-19 is reshaping the role and modes of higher education whilst moving towards a knowledge society: The case of the UAE. *Open Learning: The Journal of Open, Distance and e-Learning*, 39(1), 52-67. <https://doi.org/10.1080/02680513.2021.19305226>
- Baran, B. (2010). Experiences from the process of designing lessons with interactive whiteboard: ASSURE as a road map. *Contemporary Educational Technology*, 1(4), 367-380. <https://doi.org/10.30935/cedtech/6039>
- Borg, S. (2001). The research journal: A tool for promoting and understanding researcher development. *Language Teaching Research*, 5(2), 156-177. <https://doi.org/10.1177/136216880100500204>
- Boyadzhieva, E. (2014). Theory and practice in foreign language teaching and present. *Journal of Modern Education Review*, 4(10), 776-788. [https://doi.org/10.15341/jmer\(2155-7993\)/10.04.2014/006](https://doi.org/10.15341/jmer(2155-7993)/10.04.2014/006)
- Brown, A.H., & Green, T.D (2006). *The essentials of instructional design: Connecting fundamental principles with process and practice*. Pearson.
- Cooke, K.N. (2008). *A study of an educational blogging environment in the context of the ARCS model of motivation* [Unpublished doctoral dissertation]. University of Virginia. <https://doi.org/10.21203/rs.3.rs-1234567/v1>

[org/10.18130/V3G524](https://doi.org/10.18130/V3G524)

- Creswell, J.W. (2007). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). SAGE
- Çatar, B., & Özdilek, Z. (2022). Türkiye’de ASSURE öğretim tasarımı modeli alanında yayınlanan araştırmaların betimsel içerik analizi [A descriptive content analysis of research on the ASSURE instructional design model in Turkey]. *Fen, Matematik, Girişimcilik ve Teknoloji Eğitimi Dergisi*, 5(2), 123-144.
- Çibir, A., & Yazgan, Y. (2021). ASSURE öğretim tasarım modeline dayalı ders tasarımının ilkökul ikinci sınıfta zihinden toplama işlemindeki başarıya etkisi [The impact of ASSURE-based lesson design on second graders’ achievement in mental addition]. *Opus Uluslararası Toplum Araştırmaları Dergisi*, 18(39), 485-520. <https://doi.org/10.26466/opus.846504>
- Çetinkaya, M., & Taş, E. (2016). Web destekli ve etkinlik temelli ölçme değerlendirme materyali geliştirilmesi [Designing web-supported and activity-based assessment materials]. *Eğitim ve Öğretim Araştırmaları Dergisi*, 5(1), 21-28.
- Eliana, N., Wati, U.A., & Rahmadona, S. (2024). Leveraging the ASSURE Model for optimized information technology-based learning media. *Al-Ishlah: Jurnal Pendidikan*, 16(3). <https://doi.org/10.35445/alishlah.v16i3.5639>
- Ger, G. (2009). Tüketici araştırmalarında nitel yöntemler kullanmanın incelikleri ve zorlukları [Nuances and Challenges of using qualitative methods in consumer research]. *Tüketici ve Tüketim Araştırmaları Dergisi*, 1(1), 1–19.
- Ghavifekr, S., & Rosdy, W.A.W. (2015). Teaching and learning with technology: Effectiveness of ICT integration in schools. *International Journal of Research in Education and Science*, 1(2), 175-191. <https://doi.org/10.21890/ijres.23596>
- Giang, T.T.T., Andre, J., & Lan, H.H. (2022). Student engagement: Validating a model to unify in-class and out-of-class contexts. *SAGE Open*, 12(4). <https://doi.org/10.1177/21582440221140334>
- Gustafson, K.L., & Branch, R.M. (1997). Re-visioning models of instructional development. *Educational Technology Research and Development*, 45(3), 73-89. <https://doi.org/10.1007/BF02299731>
- Gündüzalp, C., & Yıldız, E.P. (2020). ASSURE modeli ile tasarlanmış bir dersin öğrencilerin bilgi iletişim teknolojileri kullanımına yönelik tutum ve bilgisayar kaygı düzeylerine etkisi [The effect of an ASSURE model based lesson on students’ attitudes toward ICT use and their computer anxiety levels]. *Ekev Akademi Dergisi*, 24(83), 107-137.
- Hadi, M. (2022). The foundation of curriculum renewal (reviewing from philosophical, juridic, historical, psychological, social, and cultural aspects). *Jurnal Pendidikan Indonesia: Teori, Penelitian dan Inovasi*, 2(2). <https://doi.org/10.59818/jpi.v2i2.202>
- Hancock, R.D., & Algozzine, B. (2006). *Doing case study research*. Teachers College Press.
- Heinich, R., Molenda, M., Russell, J.D., & Smaldino, S.E. (1999). *Instructional media and technologies for learning* (6th ed.). Prentice Hall.
- Heinich, R., Molenda, M., Russell, J.D., & Smaldino, S.E. (2001). *Instructional media and technologies for learning* (7th ed.). Prentice Hall.
- Hu, H., & Huang, F. (2022). Application of universal design for learning into remote English education in Australia amid COVID-19 pandemic. *International Journal on Studies in Education*, 4(1), 55-69. <https://doi.org/10.46328/ijonse.59>
- İrmiş, S., & Uludağ, A.K. (2023). The effects of blended learning activities based on the ASSURE model in teaching on students and teachers in music lessons. *International Journal of Assessment Tools in Education*, 10(2), 303-330. <https://doi.org/10.21449/ijate.1217352>
- Jones, C. (2019). Capital, neoliberalism and educational technology. *Postdigital Science and Education*, 1(2), 288–292. <https://doi.org/10.1007/s42438-019-00042-1>
- Juan, X. (2011). ASSURE model training for teachers of teaching Chinese as a second language. In *2011 International Conference on Consumer Electronics, Communications and Networks (CECNet)* (pp. 2984-2987). IEEE. <https://doi.org/10.1109/CECNET.2011.57695>

## 01

- Kang, B. (2021). How the COVID-19 pandemic is reshaping the education service. In J. Lee & S.H. Han (Eds.), *The future of service post-COVID-19 pandemic, volume 1: The ICT and evolution of work* (pp. 25–38). Springer. [https://doi.org/10.1007/978-981-33-4126-5\\_2](https://doi.org/10.1007/978-981-33-4126-5_2)
- Karadeniz, H., & Karamustafaoğlu, S. (2022). ASSURE öğretim tasarım modeline yönelik etkinlik geliştirme: Sindirim sistemi [Activity development based on the ASSURE instructional design model: The digestive system]. *Asya Studies*, 6(20), 23-36 <https://doi.org/10.31455/asya.1035839>
- Kaya, S., İnaç, H., & Çelik, H. (2020). Assure öğretim tasarımı uygulamalarının öğrencilerin akademik başarısı üzerine etkisi [The effect of ASSURE instructional design practices on students' academic achievement]. *Proceedings of the International Marmara Sciences Congress* (p. 471-477). Kocaeli: Kocaeli Üniversitesi.
- Kazancı, M.G., Altun, S., & Yabaş, D. (2020). Investigation of the effect of course design prepared according to ASSURE model principles on students. *Ulakbilge Sosyal Bilimler Dergisi*, 8(54), 1265-1276. <https://doi.org/10.7816/ulakbilge-08-54-01>
- Kim, D., & Downey, S. (2016). Examining the use of the ASSURE model by K-12 teachers. *Computers in the Schools*, 33(3), 153-168. <https://doi.org/10.1080/07380569.2016.1203208>
- Kristianti, Y., Prabawanto, S., & Suhendra, S. (2017). Critical thinking skills of students through mathematics learning with ASSURE model assisted by software Autograph. In *International Conference on Mathematics and Science Education (ICMScE)*, 875(1), 1-5. IOP Publishing.
- Lei, G. (2023). Influence of ASSURE model in enhancing educational technology. *Interactive Learning Environment*, 1-17. <https://doi.org/10.1080/10494820.2023.2172047>
- Marín, V.I., Duarte, J.M., Galvis, A.H., & Zawacki-Richter, O. (2018). Thematic analysis of the International Journal of Educational Technology in Higher Education (ETHE) between 2004 and 2017. *International Journal of Educational Technology in Higher Education*, 15(1), 1–7. <https://doi.org/10.1186/s41239-018-0089-y>
- Maican, M., & Cocoradă, E. (2021). Online foreign language learning in higher education and its correlates during the COVID-19 pandemic. *Sustainability*, 13(2), 781. <https://doi.org/10.3390/su13020781>
- Megaw, A.E. (2006). Deconstructing the Heinich, Molenda, Russell, and Smaldino instructional design model. Number Systems – Online Digital Electronics Course. (n.d). Retrieved August 25, 2018, from <http://electronics-course.com/number-systems>
- Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook*. (2nd ed.). Sage Publication.
- Ministry of National Education (MoNE) (2018). English language curriculum (Grades 1-8). <https://mufredat.meb.gov.tr/>
- Mills, G.E. (2003). *Action research: A guide for the teacher researcher*. Merrill Prentice Hall.
- Molenda, M., Pershing, J.A., & Reigeluth, C.M. (1996). Designing instructional systems. In Craig, R.L. (Ed), *The ASTD training & development handbook: A guide to human resource development* (pp. 266–293). McGraw-Hill.
- Mulya, N.T., & Putro Setyo, N.H. (2024). Technological interventions in English pedagogy: A comprehensive analysis of their influence on teaching and learning of the English language. *Al-Ishlah*, 16(4), 4690–4700. <https://doi.org/10.35445/alishlah.v16i4.4716>
- Mwanza, D.S. (2017). The eclectic approach to language teaching: Its conceptualization and misconceptions. *International Journal of Humanities Social Sciences and Education*, 4(2), 53-67. <http://dx.doi.org/10.20431/2349-0381.0402006>
- Nagy, T. (2021). Using technology for foreign language learning: The teacher's role. *Central European Journal of Education Research*, 3(2), 23-28. <https://doi.org/10.37441/cejer/2021/3/2/9347>
- Poloju, R. (2024). Revolutionizing English language teaching: The impact of technology on language learning. <https://doi.org/10.69758/gimrj/2412ivvxiip0007>

- Reigeluth, C.M. (1983). Instructional design: What is it and why is it. In C.M. Reigeluth (Ed.), *Instructional-design theories and models: An overview of their current status* (3-36). Lawrence Erlbaum Associates.
- Reiser, R.A., & Dempsey, J.V. (2008) *Trends and issues in instructional design and technology*. (2nd ed., pp. 312-321). Pearson Education.
- Rodríguez-Izquierdo, R.M. (2021). Perceptions of linguistically responsive teaching in language specialist teachers and mainstream teachers. *Porta Linguarum*, 35, 25-41. <http://dx.doi.org/10.30827/portalin.v0i35.16859>
- Russell, J.D., Sorge, D., & Brickner, D. (1994). Improving technology implementation in grades 5–12 with the ASSURE model. *Technological Horizons in Education*, 21(9), 66–70.
- Russell, J.D., & Butcher, C. (1999). Using portfolios in educational technology courses. *Journal of Technology and Teacher Education*, 7(4), 279–289.
- Sezer, B., Karaoğlu Yılmaz, F.G., & Yılmaz, R. (2013). Integrating technology into classroom: The learner-centered instructional design. *International Journal on New Trends in Education and Their Implications*, 4(4), 134-144.
- Sari, D.M., & Hasanah, M. (2022). Implementation of Canva application-based Assure model learning design in fiqh learning. *J-PAI: Jurnal Pendidikan Agama Islam*, 9(1). <https://doi.org/10.18860/jpai.v9i1.19020>
- Saputra, N., Amiruddin, A., & Saputra, M. (2021). Application of the Assure learning model in improving the learning outcomes of class IV elementary school students. *ZAHRA: Research and Tough Elementary School of Education*, 2(2), 112-122. <https://doi.org/10.37812/zahra.v2i2.198>
- Shadiev, R., & Wang, X. (2022). A review of research on technology-supported language learning and 21st century skills. *Frontiers in Psychology*, 13, 897689. <https://doi.org/10.3389/fpsyg.2022.897689>
- Shelly, G.B., Gunter, G.A., & Gunter, R.E. (2012). *Teachers discovering computers: Integrating technology in a connected world*. Cengage Learning.
- Smaldino, S., Russel, J.D., Heinich, R., & Molenda, M. (2005). *Instructional technology and media for learning*. Prentice Hall.
- Smaldino, S.E., Lowther, D.L. & Russell, J.D. (2008). *Instructional technology and media for learning*. (9th ed.). Pearson.
- Smaldino, S.E., Lowther, D.L., Russell, J.D., & Mims, C. (2015). *Instructional technology and media for learning* (10th ed.). Prentice Hall.
- Smith, P.L., & Ragan, T.J. (1999). *Instructional design*. John Wiley & Sons Inc.
- Stake, R.E. (1995). *The art of case study research*. Sage Publication.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publication.
- Storey, L. (2007). Doing interpretative phenomenological analysis. In E. Lyons & A. Coyle (Eds.). *Analysing qualitative data in psychology*. (pp. 51-64). Sage.
- Sun, L., Tang, Y., & Zuo, W. (2020). Coronavirus pushes education online. *Nature Materials*, 19(6), 687. <https://doi.org/10.1038/s41563-020-0678-8>
- Sundayana, R., Herman, T., Dahlan, J.A., & Prahmana, R.C.I. (2017). Using ASSURE learning design to develop students' mathematical communication ability. *Word Transactions on Engineering and Technology Education*, 15(3), 245-249.
- Şahbaz, E., Karabulut, H., Gökçe, H., & Kariper, İ. A. (2024). The development of an environmental educational game based on the ASSURE instructional design model: A case study in Turkey. *Education*, 3-13, 1–17. <https://doi.org/10.1080/03004279.2024.2412694>
- Şimşek, A. (2017). *Öğretim tasarımı*, (5.baskı) [Instructional design], (5th ed.). Pegem Akademi.
- Tosun, C. (2012). Yurdumuzda yabancı dil öğretme ve öğrenme sürecinde başarısızlığın nedeni yöntem mi? [Is the method the reason for failure in the process of teaching and learning foreign languages in our country?] In A. Sarıçoban & H. Öz (Eds), *Türkiye’de yabancı dil*



- eğitimde eğilim ne olmalı?* 1.Yabancı Dil Eğitim Çalıştayı Bildirileri (12-13 Kasım 2012) [What should be the trend in foreign language education in Türkiye? Proceedings of 1st Foreign Education Workshop (12-13 November 2012)] (pp. 37- 40) Hacettepe Üniversitesi Yayınları. <https://doi.org/10.13140/2.1.4171.4884>
- Whitelock, D. (2024). Innovation and adaption during the COVID-19 pandemic. *Open Learning: The Journal of Open, Distance and e-Learning*, 39(1) pp. 1-3. <https://doi.org/10.1080/02680513.2023.2293696>
- Yıldırım, A., & Şimşek, H. (2008). *Sosyal bilimlerde nitel araştırma yöntemleri*, (6.baskı). [Qualitative research methods in social sciences], (6th ed.). Seçkin Yayıncılık.
- Yin, R.K. (1984). *Case study research: Design and methods*. Sage Publication.
- Yin, R.K. (2003). *Case study research: Design and methods*. (3rd ed.). Sage Publication.
- Zai, A.G., Halawa, S., & Waruwu, R.S. (2024). Analysing the effect of using ASSURE learning model on the development of students' learning motivation. *Journal of Historical Education Studies*, 1(2), 119–127. <https://doi.org/10.61677/satmata.v1i2.175>
- Zahran, F.A. (2023). The Impact of ASSURE model-based program on EFL in-service preparatory teachers teaching skills and digital literacy skills. *International Journal of Research in Education and Science*, 9(4), 883-900. <https://doi.org/10.46328/ijres.3279>
- Zhu, Y., Pang, W., & Chen, B.B. (2023). The student engagement scale: Evidence of psychometric validity in Chinese and English language subjects from grade 4 to grade 6 in China. *Educational Psychology*, 43(2-3), 173-186. <https://doi.org/10.1080/01443410-2023-2169253>



## ChatGPT vs. DeepSeek: A comparative psychometric evaluation of AI tools in generating multiple-choice questions

Ceylan Gündeğer Kılci<sup>1\*</sup>

<sup>1</sup>Aksaray University, Faculty of Education, Department of Educational Sciences, Aksaray, Türkiye

### ARTICLE HISTORY

Received: Apr. 12, 2025

Accepted: Aug. 3, 2025

### Keywords:

ChatGPT,  
DeepSeek,  
Item generation,  
Psychometrics,  
Generalizability theory.

**Abstract:** This study examined the psychometric quality of multiple-choice questions generated by two AI tools, ChatGPT and DeepSeek, within the context of an undergraduate Educational Measurement and Evaluation course. Guided by ten learning outcomes (LOs) aligned with Bloom's Taxonomy, each tool was prompted to generate one five-option multiple-choice item per LO. Following expert review (Kendall's  $W = .58$ ); revisions were made, and the finalized test was administered to 120 students. Item analyses revealed no statistically significant differences between the two AI models regarding item difficulty, discrimination, variance, or reliability. A few items -two from ChatGPT and one from DeepSeek- had suboptimal discrimination indices. Tetrachoric correlation analyses of item pairs generated by the two AI tools for the same LO revealed that only one pair showed a non-significant association, whereas all other pairs demonstrated statistically significant and generally moderate correlations. KR-20 and split-half reliability coefficients reflected acceptable internal consistency for a classroom-based assessment, with the DeepSeek-generated half showing a slightly stronger correlation with total scores. Expert feedback indicated that while AI tools generally produced valid stems and correct answers, most revisions focused on improving distractor quality, highlighting the need for human refinement. Generalizability and Decision studies confirmed consistency in expert ratings and recommended a minimum of seven experts for reliable evaluations. In conclusion, both AI tools demonstrated the capacity to generate psychometrically comparable items, highlighting their potential to support educators and test developers in test construction. The study concludes with practical recommendations for effectively incorporating AI into test development workflows.

## 1. INTRODUCTION

In recent years, the integration of artificial intelligence (AI) into educational measurement has gained significant momentum. One of the most promising applications of AI in this field is the generation of test items, traditionally a time-consuming process that requires substantial subject-matter expertise. The advent of large language models (LLMs), such as ChatGPT (Open AI, 2023) and DeepSeek (DeepSeek AI, 2024), has opened new possibilities for automating the development of multiple-choice questions (MCQs), a widely used item format in achievement testing. Particularly, Automatic Item Generation (AIG) represents a transformative approach,

\*CONTACT: Ceylan GÜNDEĞER KILCI ✉ [cgundeger@gmail.com](mailto:cgundeger@gmail.com) 📍 Aksaray University, Faculty of Education, Department of Educational Sciences, Aksaray, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

leveraging advanced technologies for the efficient development and validation of test items (Gierl & Haladyna, 2013). AIG is defined as the process of producing MCQs using cognitive models and computer algorithms (Gierl *et al.*, 2021) and is increasingly viewed as a cost-effective and scalable solution for constructing large item banks (Gierl *et al.*, 2012).

The widespread use of MCQs in large-scale examinations underscores the growing relevance of AIG. For instance, in high-stakes assessments such as the TOEFL and GRE, AI-supported systems like e-rater and SpeechRater are used to evaluate students' writing and speaking performance (Educational Testing Service, 2005a, 2005b). Similarly, the Graduate Management Admission Test (GMAT) includes the *Analytical Writing Assessment*, which incorporates automated scoring procedures (Graduate Management Admission Council, 2009). In the Programme for International Student Assessment (PISA), AI technologies are not only used to deliver adaptive testing but also to analyze students' reasoning processes (OECD, 2023). The SAT employs AI algorithms to detect irregular testing behavior (The Princeton Review, 2024), and in China's National College Entrance Examination (Gaokao), students' essays are routinely scored by AI-based systems (Xiaoyu, 2024). Furthermore, some leading educational institutions around the world have publicly stated that future large-scale assessments may be developed entirely using AI tools (Daily Sabah, 2023; ÖSYM, 2024). These developments reflect a growing institutional trust in AI's potential to enhance—or even replace—conventional assessment design and implementation.

While MCQs allow students to interact with key concepts and enable instructors to monitor learning progress (Malik *et al.*, 2024), their development and reuse in classroom settings pose serious challenges. Among these, test security and time constraints in generating alternative but equivalent items are most notable. Test security is a critical issue since the same test items are used repeatedly across academic terms, and item exposure can threaten the integrity and consistency of the assessment. Similarly, when instructors attempt to generate alternative items aligned with the same learning outcome (LO), the process becomes increasingly complex and time-consuming. In this context, AI has the potential to offer valuable support. As Seldon and Abidoye (2018) have noted, AI applications in education may assist teachers and academics in overcoming the difficulties associated with item writing. Hence, Thorndike and Thorndike-Christ (2014) emphasized that if AI tools are used for creating exams, validity and reliability become critical concerns because they are required qualities of any test.

The primary objective of test development is to construct measurement tools that validly and reliably assess the intended construct. Numerous frameworks have been proposed in the literature to guide the test development process, each with slight variations (e.g., Downing & Haladyna, 2011; Irwing & Hughes, 2018; Turgut & Baykul, 2012). Among these, the current study adopted the sequence proposed by Crocker and Algina (1986) due to its conceptual clarity and practical simplicity. According to Crocker and Algina (1986), the test development process begins by clearly defining the intended use of the test scores, followed by the identification of specific behaviors that represent the target construct to be measured. Based on this information, a detailed table of specifications is developed. An initial pool of test items is constructed, which is then subjected to expert review and necessary revisions to enhance clarity, relevance, and content validity. After necessary revisions, a small-scale pilot study is conducted to gather preliminary data on item performance. The trial version of the test is administered to a sample that closely resembles the target population. Item statistics are then analyzed, and items that fail to meet predetermined psychometric criteria (such as adequate difficulty and discrimination indices) are eliminated. At this stage, validity and reliability studies are designed and implemented to evaluate the psychometric soundness of the final form. Finally, a scoring and interpretation guide is prepared to facilitate consistent administration and meaningful interpretation of the test results.

A foundational element in the test development process is Bloom's Taxonomy (Anderson & Krathwohl, 2001; Bloom *et al.*, 1956), which plays a central role particularly during the

identification of LOs and construction of the table of specifications. The taxonomy was developed as a framework for classifying student behaviors that reflect the intended outcomes of the educational process, and it contains six major classes: Knowledge, Comprehension, Application, Analysis, Synthesis, Evaluation (Bloom *et al.*, 1956). Since Bloom's Taxonomy offers a structured framework for articulating LOs across different levels of cognitive complexity, it serves as a key reference point when evaluating whether AI-generated test items appropriately reflect the intended educational objectives.

AI tools have the potential to significantly enhance the efficiency of test item development (Shin, 2023). Considering all these points, several critical questions emerge regarding the utility of AI in test development: Can AI tools produce high-quality multiple-choice questions that align with specific LOs? Which tools' items receive more expert suggestions for revision, and how consistent are expert evaluations? What is the optimal number of experts required to ensure acceptable generalizability? Most importantly, which AI tool demonstrates greater potential in generating psychometrically high-quality multiple-choice items suitable for classroom-based assessment contexts?

### 1.1. What is high-quality MCQ?

According to Clauser *et al.* (2006), MCQs offer the advantages of assessing broad content in a short period of time, being resistant to subjective grading, and providing accurate and efficient scoring. However, the quality of these items—particularly those generated by AI—must still be evaluated against fundamental measurement criteria. Test items must align with learning objectives (LOs) and demonstrate acceptable levels of item difficulty, discrimination, and content validity (Haladyna & Rodriguez, 2013). Items that are considered psychometrically high-quality are expected to exhibit strong discriminative power. The item discrimination index, one of the most critical indicators in test development, can be calculated within the framework of Classical Test Theory (CTT), using either the biserial correlation coefficient or the upper-lower group method. A discrimination index of .30 or higher is generally accepted as sufficient, indicating that the item effectively differentiates between higher- and lower-performing examinees and is suitable for inclusion in the final version of the test (Crocker & Algina, 1986).

In addition to discrimination, items are expected to demonstrate high variance and item-level reliability (Atılgan, 2009; Tan, 2014). In multiple-choice tests, the maximum item variance is .25, which occurs when the item difficulty index is .50. As the variance (or standard deviation) of an item increases, its ability to reveal individual differences in the trait being measured also increases (Baykul, 2000). Item reliability, which is positively associated with both item variance and discrimination, increases proportionally with these values. The reliability of a test is positively influenced by the reliability of its individual items; as the reliability of individual items increases, so does the reliability of the entire test (Atılgan, 2009). Expert review also plays a critical role in establishing content validity. To minimize potential errors and ensure content appropriateness, it is recommended that multiple-choice test items generated by AI tools undergo thorough review prior to their implementation (Kıyak & Emekli, 2024). As Crocker and Algina (1986) emphasized, the consistency among expert judgments should be carefully considered during the test development process. To determine the optimal number of experts required for reliable content validation, researchers may draw on Generalizability Theory, which provides a framework for estimating the reliability of expert evaluations across various measurement scenarios.

### 1.2. Generalizability (G) Theory

G Theory is a statistical framework used to examine the reliability of performance assessments (Shavelson & Webb, 1991). It provides a comprehensive conceptual model and a rigorous statistical approach for a wide range of measurement issues. Often regarded as an extension of CTT, G Theory incorporates analysis of variance (ANOVA) techniques. While CTT assumes

that observed scores are composed of a true score ( $T$ ) and random error ( $E$ ), it does not allow for the identification of distinct sources of error variance. In contrast, G Theory enables researchers to examine multiple sources of measurement error, collectively denoted as ( $E$ ) in CTT, within a single analytic framework (Brennan, 2001; Shavelson & Webb, 1991). Moreover, G Theory introduces two reliability coefficients: the generalizability coefficient ( $G$ ) and the Phi ( $\Phi$ ) coefficient, which support both relative (norm-referenced) and absolute (criterion-referenced) interpretations (Shavelson & Webb, 1991). Through Decision ( $D$ ) studies, researchers can also estimate how variations in the number of items, raters/experts, or other measurement facets may influence the reliability and validity of test scores (Crocker & Algina, 1986). Essentially, G Theory distinguishes between Generalizability ( $G$ ) studies and Decision ( $D$ ) studies: the former estimates sources of measurement error, while the latter uses these estimates to simulate optimal testing conditions that minimize error for specific assessment purposes.  $D$  studies thus inform practical decisions regarding the design and administration of assessments in order to achieve the desired level of reliability (Brennan, 2001; Crocker & Algina, 1986; Shavelson & Webb, 1991). Despite the recognized importance of expert evaluations in the test development process, the literature lacks empirical studies that utilize G Theory to assess the reliability of expert ratings or to identify the minimum number of raters/experts needed to ensure acceptable generalizability. To date, only one study in the literature has examined DeepSeek in the context of EFL writing evaluation using G Theory. The findings revealed that DeepSeek-V3 consistently demonstrated lower scoring reliability than both DeepSeek-R1 and human raters. However, DeepSeek-R1 outperformed human raters by consistently yielding higher reliability coefficients (Gao *et al.*, 2025).

A growing body of literature highlights the diverse applications of AI tools (particularly ChatGPT) in educational contexts. These applications can be broadly categorized into two main domains: *teaching preparation* and *assessment* (Lo, 2023). The *teaching preparation* category encompasses generating lesson materials, providing suggestions, and performing language translation, whereas the *assessment* category includes generating assessment tasks and evaluating student performance. Given the current study's focus on analyzing the psychometric properties of MCQs generated by two distinct AI tools, it aligns specifically with the “assessment tasks” subdomain under the broader “assessment” category. Research within this domain indicates that AI-generated assessment content often requires structured guidance to minimize the likelihood of inaccurate or low-quality outputs (Urhan *et al.*, 2024; Wardat *et al.*, 2023). A well-designed guidance can assist teachers in developing high-quality items and tests for classroom-based assessment contexts (Kıyak & Emekli, 2024). In contrast, studies conducted without appropriate guidance suggest that AI tools are not sufficient (Ngo *et al.*, 2024). These considerations raise an important question: How can AI tools be leveraged most effectively in classroom-based assessment to develop valid and reliable tests? To address this question, the current study was designed to compare two prominent LLMs, ChatGPT and DeepSeek, in generating MCQs aligned with ten LOs in the Educational Measurement and Evaluation course. A standardized prompt was developed, including only essential elements: Course, grade level, LO and corresponding Bloom's Taxonomy level, the desired number of response options and correct answers. Focusing on the Educational Measurement and Evaluation course, the study examined how ChatGPT and DeepSeek performed when generating items aligned with ten specific LOs using this simple prompt structure. The psychometric quality of the items was subsequently evaluated. A similar comparative study by Rycroft-Smith and Macey (2024) assessed MCQs generated by Copilot, Claude, and ChatGPT 3.5 on the topic of the area of rectangles. The results revealed notable differences among the tools in terms of item length (word count), mathematical accuracy, and capacity to elicit student understanding. Claude produced the highest proportion of mathematically accurate items, whereas Copilot generated the shortest items and ChatGPT the longest. In another study, Ünal *et al.* (2025) evaluated ChatGPT's item-writing performance in the domain of number sense.



Of the 27 generated items, only three were found to lack acceptable item-level statistics. The remaining 24 items formed the final version of the test, which yielded a Cronbach's Alpha of .88, indicating a high level of internal consistency. The test was also determined to have moderate difficulty and sufficient discrimination.

Although the literature on AIG is growing, empirical studies specifically examining widely used AI tools such as ChatGPT and DeepSeek—and evaluating the psychometric properties of the items they generate—remain limited (Alafnan, 2025; Kanik, 2024; Leslie & Gierl, 2023; Wang & Heung, 2025; Zhai, 2025). This study introduces an empirical framework designed to investigate how effective and high-quality test items can be developed by teachers or academics using AI tools in their most basic and accessible form for classroom assessment purposes. Furthermore, there is a notable lack of comparative research on different LLMs within the context of educational measurement. Given that each model may differ in linguistic style, response consistency, and alignment with domain-specific content (Liu *et al.*, 2023), directly comparing their capacity to generate valid and reliable test items can offer valuable insights into their respective strengths and limitations. In a recent study conducted by Yogesh *et al.* (2025), ChatGPT and DeepSeek were compared across several criteria, including accuracy, usability, response consistency, domain-specific knowledge, and computational efficiency. The findings indicated that ChatGPT demonstrated versatile performance in general-purpose tasks such as conversational abilities, creativity, and content generation. In contrast, DeepSeek stood out in specialized domains by providing precise and domain-specific responses, particularly in areas such as technical problem-solving and scientific inquiry. This current study aims to explore the potential of AI tools in achievement test development through a comparative analysis of MCQs generated by two prominent LLMs: ChatGPT and DeepSeek. Specifically, the study examines the psychometric quality of the items produced by each model, focusing on both item-level and test-level statistics, as well as expert evaluation via G Theory. In doing so, the research contributes to the emerging literature on AI-assisted educational assessment and provides practical implications for test developers, educators, psychometricians, and policymakers seeking to integrate AI technologies into both classroom-based and large-scale assessment practices.

### 1.3. Purpose of the Study

The primary aim of this study is to compare the psychometric properties of multiple-choice items developed using ChatGPT and DeepSeek. The study seeks to evaluate the quality of AI-generated items intended for classroom-based use, with a particular focus on content validity, item difficulty, discrimination, and reliability. To achieve this aim, the study addresses the following research questions:

1. How are the item-level statistics (item difficulty, discrimination, variance, and reliability) of the multiple-choice questions generated by ChatGPT and DeepSeek?
2. Do the item-level statistics of the items generated by the two AI tools show a statistically significant difference?
3. What is the relationship between the item pairs generated by the two AI tools for the same LO?
4. How are the KR-20 and split-half reliability coefficients of both the trial and final forms of the test?
5. Do the average scores from the half-tests generated by ChatGPT and DeepSeek differ significantly?
6. What is the relationship between students' total test scores and their scores on each half-test?
7. Which items generated by ChatGPT and DeepSeek were identified by field experts as requiring revision?
8. How are the variance components associated with expert ratings, and according to the D study, what is the optimal number of raters needed to ensure acceptable generalizability?



## 2. METHOD

### 2.1. Study Design

The primary objective of this study was to examine the psychometric properties of multiple-choice test items generated by ChatGPT and DeepSeek and to compare their item-level statistics. A quantitative research methodology was employed, utilizing a design that combined both descriptive and correlational approaches. While descriptive research presents findings as they are, correlational research aims to analyze the relationships between two or more variables (Fraenkel *et al.*, 2012).

### 2.2. Participants and Data Collection

The trial version of the test was administered to a convenience sample of 120 undergraduate students enrolled in the Faculty of Education at Aksaray University during the spring semester of the 2024-2025 academic year. This sample size aligns with various recommendations in the literature regarding appropriate participant numbers for test development. For instance, Özçelik (1992) suggests that sample sizes may range from 120 to 400, while Crocker and Algina (1986) recommend that a sample of 100 to 200 participants is sufficient. Participants responded to the 20-item multiple-choice test under standard classroom conditions. No time limit was imposed, and participation was voluntary and anonymous. The necessary ethical approval for this study was obtained from the institutional ethics committee (Approval Number: 2025/03, Aksaray University Human Research Ethics Committee). Since the aim of this study is to examine the MCQs for the undergraduate-level Educational Measurement and Evaluation course with the assistance of AI tools, ChatGPT and DeepSeek, and to provide evidence for the test's validity and reliability, the remaining stages of the test development process were followed in the subsequent phases of the study, including item-level and test-level statistics.

### 2.3. Item Generation Process

In this study, the test development stages outlined by Crocker and Algina (1986) were followed. The first step involved defining the intended use of the test. Taking into account the researcher's area of specialization, the test was developed to measure student achievement in the undergraduate-level Educational Measurement and Evaluation course offered in the Faculty of Education. To identify the attributes to be measured, ten critical LOs were determined. These LOs were selected specifically for the target course and were deemed suitable for assessment through a multiple-choice test. All LOs correspond to the Comprehension level of Bloom's Taxonomy (Bloom *et al.*, 1956). This decision was made to avoid introducing additional sources of variability that might arise from using items across different cognitive levels of Bloom's taxonomy.

Ten LOs served as the foundation for constructing an initial item pool using AI tools, ChatGPT (4.0) and DeepSeek (Deep Think R1). Using only a standardized and basic prompt structure (Appendix 1), both ChatGPT and DeepSeek were simultaneously asked to generate multiple-choice questions aligned with each LO, without receiving any additional training or manual intervention. Each prompt included the specified LO, its corresponding taxonomic level, and a brief guideline containing faculty, course, and instructional context details. Based on this information, the AI tools were instructed to generate one multiple-choice item consisting of five answer options with one correct response and four plausible distractors. The generated items were reviewed by the researcher to ensure compliance with the prompt specifications. If an item failed to meet the specified criteria, did not adequately align with the intended LOs, or included scientific inaccuracies, the AI tool was requested to produce a new item, limited to a maximum of three iterations. In total, 20 items (10 from each AI model) were selected for inclusion in the trial version of the test.

## 2.4. Expert Review and Content Validation

Items were reviewed by three field experts for content relevance and clarity. The experts consisted of academics who had completed their doctoral degrees, were employed at public or private universities, and had been teaching undergraduate-level Educational Measurement and Evaluation courses for at least two years. They were asked to evaluate each item in terms of its alignment with the intended LOs using a three-point rating scale: *1 – Not appropriate*, *2 – Partially appropriate*, and *3 – Appropriate*. Additionally, they were invited to provide written suggestions for the improvement of each item. Moreover, the MCQs were refined in accordance with the qualitative feedback provided by the experts, ensuring improved alignment with the intended LOs and enhanced clarity. To assess the consistency of expert judgments, Kendall's coefficient of concordance  $W$  (Kendall & Smith, 1939), which is an index of inter-rater reliability of ordinal data, was computed as .58 (Kendall chi-squared = 33.132,  $df = 19$ , subjects = 20, raters = 3,  $p = .023$ ), indicating a statistically significant level of moderate agreement (Landis & Koch, 1977). Alongside the evaluation of inter-rater agreement, the test was revised in accordance with the feedback and suggestions offered by the experts regarding individual items. A small pilot study was carried out with a sample of three former students, during which the items were evaluated for clarity. The results indicated that the items were comprehensible and did not require another revision, and thus, the test was considered ready for the trial administration.

## 2.5. Data Analysis

To answer the first two research questions (RQs) of the study, item-level statistics were calculated, including item difficulty, item discrimination (item-total biserial correlation), item variance, and item reliability. The item difficulty index ( $p$ ) refers to the proportion of students who answered the item correctly. The item discrimination indices were calculated based on the relationship between the item score and the total test score using the biserial correlation coefficient, since item scores are dichotomous and total scores are continuous. Item variance was calculated using the formula  $p \times (1 - p)$ , and item reliability was estimated by multiplying the item's standard deviation by the item discrimination index. Descriptive statistics of the item-level statistics were also calculated, and Mann-Whitney U tests were conducted to examine whether the obtained item-level statistics differed based on the AI tool used. To address RQ3, the relationships between the item pairs generated by ChatGPT and DeepSeek for the same LO were examined by calculating tetrachoric correlation coefficients and their corresponding  $p$  values.

At the test level, the KR-20 reliability coefficient and split-half method with Spearman-Brown correction were estimated for RQ4. Split-half reliability is considered appropriate for achievement tests composed of items that vary in terms of targeted behaviors (content domains), provided that the items assigned to each half exhibit similar difficulty levels (Baykul, 2000). In this study, items generated by ChatGPT and DeepSeek were matched for each LO represented in the test. In RQ 2, it was observed that the items assigned to the two halves had similar difficulty levels and that the item-level statistics did not differ significantly. Therefore, split-half reliability was calculated by assigning the items generated by ChatGPT to one half and those produced by DeepSeek to the other. In addition, KR-20 was used to estimate the internal consistency of the test scores. While internal reliability coefficients of .90 or higher are typically expected in high-stakes standardized testing, lower values are considered acceptable in classroom-based assessments (Rudner & Schafer, 2002). Although Murphy and Davidshofer (1991) emphasized that a minimum internal consistency coefficient of .75 is required for classroom-based achievement tests, Rudner and Schafer (2002) noted that teacher-made assessments can still be considered acceptable with reliability coefficients as low as .50 to .60.

In response to RQ5 and RQ6, paired samples t-tests were conducted to examine whether the half-test scores differed based on the AI tool used. The relationships between the test halves

and the total test scores were examined using Pearson's correlation coefficient, as the test scores showed normal distribution. Although the literature presents varying threshold values for interpreting the strength of correlation coefficients (e.g., Dancey & Reidy, 2017; Schober *et al.*, 2018), this study adopted the classification proposed by Ratner (2009), wherein values between .00 and .30 are considered weak, values between .30 and .70 are regarded as moderate, and values between .70 and 1.00 are interpreted as strong correlations. The comparison of the correlation coefficients was conducted using Meng, Rosenthal, and Rubin's (1992)  $z$  test for overlapping dependent correlations. All analyses were conducted in R (R Core Team, 2023) with the packages *psych* (Revelle, 2023), *CTT* (Willse, 2014), *dplyr* (Wickham *et al.*, 2023), *correlation* (Makowski *et al.*, 2022), *cocor* (Diedenhofen & Musch, 2015), *sirt* (Robitzsch, 2024), and *readxl* (Wickham & Bryan, 2023). To assess the reliability of expert ratings, G Theory was employed. Using a fully crossed ( $I \times E$ ) design, in which every item ( $I$ ) was evaluated by every expert ( $E$ ), variance components were calculated via a G study, while the D study simulated different rater numbers to identify the optimal panel size. Shavelson and Webb (1991) emphasized that both the  $G$  and Phi ( $\Phi$ ) coefficients should be at least .80, indicating high reliability.  $G$  and  $D$  studies were performed using EduG (Society & Group, 2010) software, which applies G Theory while providing a simple and easy-to-use display (Clauser, 2008), and the interpretation of the findings was conducted in light of these benchmarks.

### 3. RESULTS

#### 3.1. Item-Level Statistics of the AI-Generated Test Items

To explore RQ1, item difficulty ( $p$ ), item-total biserial correlation (item discrimination:  $r$ ), item variance ( $S^2$ ), and item reliability ( $IR$ ) were calculated. The items developed by the AI tools, ChatGPT and DeepSeek, were aligned with the corresponding LOs to allow for a direct comparison across tools. The results of this analysis are presented in Table 1.

**Table 1.** Item-level statistics.

LO	ChatGPT					DeepSeek				
	Item	$p$	$r$	$S^2$	$IR$	Item	$p$	$r$	$S^2$	$IR$
LO1	Item 1	.78	.48	.17	.2	Item 2	.63	.51	.23	.24
LO2	Item 4	.82	.5	.15	.19	Item 8	.6	.36	.24	.18
LO3	Item 7	.86	.3	.12	.1	Item 5	.57	.48	.25	.24
LO4	Item 10	.84	.46	.13	.17	Item 11	.81	.37	.15	.14
LO5	Item 12	.9	.4	.09	.12	Item 9	.82	.33	.15	.13
LO6	Item 14	.89	.14	.1	.04	Item 3	.9	.27	.09	.08
LO7	Item 18	.57	.38	.25	.19	Item 15	.74	.46	.19	.2
LO8	Item 19	.78	.24	.17	.1	Item 17	.41	.38	.24	.19
LO9	Item 16	.9	.34	.09	.1	Item 13	.82	.34	.15	.13
LO10	Item 6	.48	.34	.25	.17	Item 20	.58	.51	.24	.25

LO: Learning outcome;  $p$ : Item difficulty;  $r$ : Item-Total biserial correlation;  $S^2$ : Item variance;  $IR$  (Item reliability);  $r \times SD$

A general examination of the item-level statistics in Table 1 reveals that 17 out of the 20 test items have discrimination indices equal to or greater than .30, indicating acceptable item discrimination (Crocker & Algina, 1986). Only three items fall below this threshold: Item 14 and Item 19, both generated by ChatGPT, and Item 3, generated by DeepSeek. In terms of item difficulty, values range from .41 to .90, suggesting that the majority of the items fall within the moderate to easy difficulty level. Additionally, item variances vary considerably across items; while some items (e.g., Item 12, Item 16, and Item 3) exhibit very low variance, others (e.g., Item 18, Item 6, and Item 5) display the highest variance levels within the test.

### 3.2. ChatGPT and DeepSeek Comparisons Regarding Item-Level Statistics

In RQ2, both descriptive statistics and a Mann-Whitney U test were employed to examine differences in item-level statistics. Table 2 presents a summary of the descriptive statistics for item difficulty, discrimination, variance, and reliability. According to Table 2, for the items generated by ChatGPT, the mode, median, and mean values of difficulty, discrimination, variance, and reliability statistics were found to be relatively close to each other.

**Table 2.** Descriptive statistics of the item-level statistics.

	ChatGPT ( $n = 10$ )				DeepSeek ( $n = 10$ )			
	$p$	$r$	$S^2$	$IR$	$p$	$r$	$S^2$	$IR$
$M$	.78	.36	.15	.14	.69	.4	.19	.18
Median	.83	.36	.14	.15	.69	.38	.21	.19
Mode	.78	.34	.17	.1	.82	.51	.24	.24
$SD$	0.14	0.11	0.06	0.05	0.15	0.08	0.05	0.06
Skewness	-1.48	-0.63	0.82	-0.48	-0.35	0.08	-0.65	-0.27
$SE$	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Kurtosis	1.2	0.08	-0.43	-0.82	-0.69	-1.3	-0.81	-0.98
$SE$	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33
Minimum	.48	.14	.09	.04	.41	.27	.09	.08
Maximum	.90	.50	.25	.20	.90	.51	.25	.25

When standard deviation values of the items generated by ChatGPT are considered, in Table 2, the most heterogeneous statistics were item difficulty, discrimination, variance, and reliability. According to skewness and kurtosis coefficients, all item-level statistics (except for the difficulty index) exhibited characteristics of the normal distribution. The difficulty indices of items generated by ChatGPT ranged between .48 and .90; discrimination indices ranged between .14 and .50; variance values ranged between .09 and .25; and reliability ranged between .04 and .20. Overall, the items produced by ChatGPT can be classified as having high difficulty indices, though they are considered easy. Their discrimination indices clustered around .36, though in some cases they fell below the .30 threshold. The average item variance was approximately .15, with some items reaching the maximum possible variance, thus contributing significantly to the overall test score.

In Table 2, the mode, median, and mean values of item difficulty, discrimination, variance, and reliability for the items generated by DeepSeek appear to be less consistent compared to those of ChatGPT. Based on standard deviation values, the most heterogeneous measures were item difficulty, discrimination, reliability, and variance. According to the skewness and kurtosis coefficients, these item-level statistics exhibit characteristics of a normal distribution, except for a slightly platykurtic distribution in the discrimination index. The difficulty indices of DeepSeek-generated items ranged from .41 to .90; discrimination indices ranged from .27 to .51; variance values, similar to those of ChatGPT, ranged from .09 to .25; and item reliability coefficients ranged from .08 to .25. Compared to ChatGPT, the items produced by DeepSeek were slightly more difficult, yet still within the moderate difficulty range. The discrimination indices clustered around .40, though some fell as low as .27. Item variances averaged around .20 and reached their maximum possible value, indicating strong contributions to the total test score.

According to Table 2 above, both ChatGPT and DeepSeek generally produced items within an acceptable range of discrimination, while ChatGPT-generated items tended to be easier. In contrast, items produced by DeepSeek demonstrated higher average reliability, variance, and discrimination indices. Additionally, both tools yielded items with comparable maximum

variance, indicating strong individual item contributions to total test scores. Given the small sample size ( $n = 10$  items per AI tool), a non-parametric Mann-Whitney U test was conducted to determine whether the observed differences were statistically significant. According to the analysis results, item difficulty (Mann-Whitney  $U = 31.5$ ,  $z = -1.4$ ,  $p = .16$ ), discrimination (Mann-Whitney  $U = 39.5$ ,  $z = -0.8$ ,  $p = .43$ ), variance (Mann-Whitney  $U = 33.5$ ,  $z = -1.26$ ,  $p = .21$ ), and reliability (Mann-Whitney  $U = 29.5$ ,  $z = -1.56$ ,  $p = .12$ ) statistics did not differ significantly based on the AI tool used for item generation ( $p > .05$ ). In other words, the differences in item-level statistics presented in Tables 1 and 2 were not statistically significant. Both ChatGPT and DeepSeek produced items with comparable levels of difficulty, discrimination, variance, and reliability.

### 3.3. Relationships Between the Items Generated by the Two AI Tools for the Same LO

To examine the relationships between the items, tetrachoric correlation coefficients were computed given the binary nature of the item responses, and corresponding  $p$  values were also obtained. The full set of inter-item correlations for all test items can be found in Appendix 2. Table 3 presents the correlations between item pairs generated by the two AI tools for the same LO.

**Table 3.** Tetrachoric correlation coefficients between item pairs.

Learning Outcome (LO)	ChatGPT items	DeepSeek items	$r$	$p$ -value
LO1	Item 1	Item 2	.52	.000
LO2	Item 4	Item 8	.58	.000
LO3	Item 7	Item 5	.25	.006
LO4	Item 10	Item 11	.49	.000
LO5	Item 12	Item 9	.38	.000
LO6	Item 14	Item 3	-.09	.310
LO7	Item 18	Item 15	.35	.000
LO8	Item 19	Item 17	.19	.034
LO9	Item 16	Item 13	.27	.003
LO10	Item 6	Item 20	.35	.000

As shown in Table 3, the correlation coefficients between item pairs generated for the same LO were statistically significant, with the exception of the pair corresponding to LO6 (*Lists possible sources of random error based on a given scenario*). The items representing LO6—Item 14 (ChatGPT) and Item 3 (DeepSeek)—exhibited no significant relationship. As indicated in Table 1, both of these items demonstrated notably low discrimination indices, which may explain the lack of correlation between them, possibly due to their insufficient ability to differentiate among students.

In Table 3, the correlation coefficients calculated for item pairs generated for the same LO generally range from .19 to .58. The weakest relationship was observed between the items corresponding to LO8 (*Selects the appropriate reliability estimation technique based on a given scenario*), with a correlation coefficient of  $r = .19$  ( $p = .034$ ). A weak but statistically significant association was found between Item 19 and Item 17 at the .05 significance level. As noted in Table 1, similar to Item 14 and Item 3, Item 19 also falls below the expected threshold for discrimination.

For the item pairs generated for LO3 (*Differentiates among nominal, ordinal, interval, and ratio scales based on a given scenario*) and LO9 (*Compares different types of tests in terms of guessing probability*), the correlation coefficients were calculated as .25 and .27, respectively ( $p < .01$ ). These results indicate a weak to moderate association between the item pairs for these two LOs at the .01 significance level. The item pairs corresponding to the remaining six LOs exhibited moderate correlations, all statistically significant at the .01 level.



### 3.4. KR-20 and Split-Half Reliability

To address RQ4, the KR-20 internal consistency coefficient and split-half reliability were calculated based on the binary nature of the 20-item trial and 17-item final tests. The KR-20 coefficient calculated for the trial test, which had a mean score of 14.68 ( $\pm 3.18$ ), was .69. In the split-half reliability analysis, when the test was split according to the source of item generation tools (ChatGPT vs. DeepSeek), the Spearman-Brown corrected reliability was calculated as .73. [Appendix 3](#) presents the reliability-related statistics calculated to identify which items contributed most and least to the overall reliability of the trial test. According to [Appendix 3](#), removing Item 14 and Item 19 resulted in a slight increase in the overall reliability coefficient. In contrast, removing any of the other items led to a decrease in reliability.

[Table 1](#) shows that two items (Item 14 and Item 19) generated by ChatGPT and one item (Item 3) generated by DeepSeek exhibited discrimination indices lower than the acceptable threshold of .30 (Crocker & Algina, 1986). The items with low discrimination indices did not substantially alter the KR-20 coefficient when removed. However, the split-half reliability improved in this scenario: When the test, using the remaining 17 items (final test), was split based on the source of the items (ChatGPT vs. DeepSeek), the Spearman-Brown corrected coefficient reached .75, with a mean score of 12.11 ( $\pm 3.00$ ). This finding indicates that the final version of the test, consisting of 17 items, may achieve an acceptable level of internal consistency reliability for a teacher-made classroom assessment (Murphy & Davidshofer, 1991).

### 3.5. Average Scores from the ChatGPT and DeepSeek Half-Tests

Descriptive statistics of the student scores obtained from the ChatGPT Half Test, the DeepSeek Half Test, and the Total Test are presented in [Table 4](#) below.

**Table 4.** Descriptive statistics of the half tests and the total test scores.

	ChatGPT Half Test	DeepSeek Half Test	Total Test
<i>N</i>	120	120	120
<i>M</i>	6.13	5.98	12.11
Median	6.0	6.0	12.0
Mode	6.0	6.0	12.0
<i>SD</i>	1.47	1.88	3.00
Skewness	-0.89	-0.64	-0.81
<i>SE</i>	0.22	0.22	0.22
Kurtosis	0.65	0.34	0.71
<i>SE</i>	0.44	0.44	0.44
Minimum	2.0	1.0	3.0
Maximum	8.0	9.0	17.0
Range	6.0	8.0	14.0

According to [Table 4](#), the score distributions for both half tests and the total test show normal distributions. The scores from both the ChatGPT and DeepSeek halves are clustered around a value of approximately 6. However, based on the standard deviation, kurtosis, and range statistics, the DeepSeek Half Test yielded a more heterogeneous distribution. This suggests that it was more effective in capturing individual differences among students compared to the ChatGPT Half Test.

As shown in [Table 4](#), the mean student scores obtained from the ChatGPT Half Test and the DeepSeek Half Test differ. The mean score for the ChatGPT Half Test is slightly higher than that of the DeepSeek Half Test. In addition, the skewness coefficient indicates that the

distribution of ChatGPT scores is more negatively skewed than DeepSeek scores. This finding suggests that the ChatGPT Half Test may have been relatively easier, resulting in higher student performance. To determine whether the difference in average scores between the two halves is statistically significant, a paired samples t-test was conducted, given that the score distributions met the assumption of normality. According to the t-test results, there is no statistically significant difference between the average scores of the ChatGPT Half Test and the DeepSeek Half Test ( $t = 1.13$ ,  $df = 119$ ,  $p = .26$ ). This suggests that student performance did not differ significantly between the test halves composed of items generated by ChatGPT and DeepSeek. This finding can be interpreted as evidence that the items produced by the two AI tools were of comparable difficulty levels.

### 3.6. Relationship Between Students' Total Test and Half-Test Scores

The relationships between the half-test scores and the total test scores were examined using Pearson's correlation coefficient. The results showed that all correlation coefficients were statistically significant at the .01 level. According to the threshold values suggested by Ratner (2009), correlation coefficients ranging from .00 to .30 are interpreted as weak, those between .30 and .70 as moderate, and those between .70 and 1.00 as strong. A moderate correlation ( $r = .60$ ,  $p < .01$ ) was observed between the two halves generated by ChatGPT and DeepSeek. In contrast, both half-tests showed strong correlations with the total test score. Notably, the correlation between the DeepSeek Half Test and the total test score ( $r = .92$ ,  $p < .01$ ) was slightly higher than the correlation between the ChatGPT Half Test and the total test score ( $r = .87$ ,  $p < .01$ ). In order to examine whether the difference between the two dependent correlations was statistically significant, the z test for comparing two overlapping correlations in dependent samples was conducted, as proposed by Meng *et al.* (1992). The results indicated a statistically significant difference between the correlation of test scores with the ChatGPT Half Test and the DeepSeek Half Test ( $z = -3.10$ ,  $p = .002$ ).

### 3.7. Items Identified by Field Experts as Requiring Revision

To examine RQ7, the evaluations and revision suggestions of field experts regarding the items generated by ChatGPT and DeepSeek for the ten LOs were taken into account. As shown in [Appendix 4](#), Items 14, 15, 17, and 19 were identified by experts as requiring revision, while the remaining items were deemed appropriate. Among these, Items 14 and 19 were generated by ChatGPT, whereas Items 15 and 17 were produced by DeepSeek, indicating an equal distribution of revision requests across the AI tools. Item 14, developed for LO6 (*Lists possible sources of random error based on a given scenario*), received feedback from two experts suggesting revisions to the correct answer. Based on the suggestions, the phrase “whether the test is valid” was revised to “using a test that is not valid.” In Item 17, generated by DeepSeek for LO7 (*Interprets a correlation matrix involving five different variables*), all the answer options initially reflected only positive or negative correlations. In line with the feedback from two experts, the response options were revised to reflect varying levels of correlation strength (weak, moderate, and strong). Both items generated for LO8 (*Selects the appropriate reliability estimation technique based on a given scenario*) received revision suggestions from experts. For Item 17, one expert recommended modifying the item stem, while another provided suggestions for revising one distractor. For Item 19, only one expert suggested a revision for a single distractor. Accordingly, among the revision suggestions received, one pertained to the reconstruction of the correct answer, another involved restructuring the item stem, while the remaining suggestions focused on the distractors. That only four out of twenty items were recommended for revision indicates a generally acceptable level of alignment between the items and the intended LOs. Moreover, the equal distribution of revision suggestions across ChatGPT and DeepSeek suggests that both tools performed comparably in generating content-appropriate items.

### 3.8. Variance Components of Expert Ratings, and Optimal Number of Raters

To estimate the variance components and variance percentages of expert judgments (see [Appendix 4](#)) regarding the 20-item test, a crossed two-facet design ( $I \times E$ ) was applied, where 20 items ( $I$ ) were evaluated by 3 experts ( $E$ ). [Table 5](#) presents the estimated variance components and the percentage of total variance explained by the main effects of items ( $I$ ), experts ( $E$ ), and the interaction term ( $I \times E$ ).

**Table 5.** The estimated variance components of the expert ratings and their percentages.

Source	<i>SS</i>	<i>df</i>	<i>MS</i>	Variance	%	<i>SE</i>
Item ( <i>I</i> )	3.52	19	0.19	0.04	37.2	0.02
Expert ( <i>E</i> )	0.13	2	0.07	0.00	0.0	0.00
<i>I</i> × <i>E</i> , <i>e</i>	2.53	38	0.07	0.07	62.8	0.01
Total	6.18	59			100	

A variance component close to zero for the main effect of experts indicates that the experts assigned similar scores to the items (Shavelson & Webb, 1991). In [Table 5](#), the variance ratio for the main effect of experts ( $E$ ) was zero, suggesting a high level of consistency among the expert raters in their evaluations. The variance ratio explained by the main effect of items ( $I$ ) was found to be 37.2%. A relatively low percentage of item-related variance may indicate that differences between items (the object of measurement) were not sufficiently captured or that the items are measuring similar characteristics. The estimated variance proportions for the item-by-expert interaction ( $I \times E$ ,  $e$ ) were 62.8%. This finding may indicate that the experts perceived the items differently or reflect a substantial amount of error variance.

According to Shavelson and Webb (1991), both the  $G$  and  $\Phi$  coefficients should be at least .80, indicating high reliability. [Table 6](#) presents the  $G$  and  $\Phi$  ( $\Phi$ ) coefficients estimated through the D study for different scenarios in which the number of experts was set to 3, 5, 7, 10, 13, and 15. The findings reveal that increasing the number of raters leads to a notable improvement in the reliability of the evaluations. Based on these results, it can be concluded that, when developing a test using items generated by AI tools, a minimum of seven expert raters should be consulted to ensure sufficient content validity and appropriateness.

**Table 6.** Estimated  $G$  and  $\Phi$  coefficients for varying numbers of experts.

Number of experts	3		5		7		10		13		15	
	$G$	$\Phi$	$G$	$\Phi$	$G$	$\Phi$	$G$	$\Phi$	$G$	$\Phi$	$G$	$\Phi$
Coefficient	.64	.64	.75	.75	.81	.81	.86	.86	.89	.89	.90	.90
SEM	0.15	0.15	0.12	0.12	0.10	0.10	0.08	0.08	0.07	0.07	0.07	0.07

## 4. DISCUSSION and CONCLUSION

The primary aim of this study was to examine and assess the psychometric quality of multiple-choice questions generated by two widely used AI tools, ChatGPT and DeepSeek, for the undergraduate-level Educational Measurement and Evaluation course. The findings of the study reveal several key insights regarding the item-level and test-level performance of both tools and their potential for use in classroom-based assessment settings. First, both AI models generated items that fell within an acceptable range of item difficulty. This suggests that both tools are capable of producing items suitable for differentiating between students with varying achievement levels. Second, with respect to item discrimination, 17 out of the 20 items demonstrated acceptable levels ( $r > .30$ ), indicating that most items were effective in distinguishing between high- and low-performing students. Notably, two items (Item 14 and Item 19) from ChatGPT and one (Item 3) from DeepSeek fell below this threshold. Items 14 and 19 were also revised based on field expert feedback. However, despite these revisions, both

items did not demonstrate adequate levels of discrimination. This finding suggests that while expert feedback is crucial for ensuring content relevance and clarity, it does not necessarily guarantee improvements in the psychometric quality of an item. In other words, expert judgment may confirm an item's alignment with LO or content validity, but empirical data are essential to evaluate how well an item functions in distinguishing between students of varying ability levels. Therefore, content-based and statistical evidence should be used in a complementary manner to support item quality throughout the test development process. Third, item variance and reliability were generally higher for items generated by DeepSeek, suggesting a stronger contribution to the overall test's reliability in terms of internal consistency. However, item-level statistics between the two sets of items generated by ChatGPT and DeepSeek did not differ statistically, indicating comparable performance overall.

Tetrachoric correlation coefficients between the item pairs generated by ChatGPT and DeepSeek for the same LO revealed noteworthy patterns. Among the ten LOs, item pairs corresponding to six demonstrated moderate and statistically significant correlations ( $p < .01$ ), indicating consistent alignment between the two AI tools for the majority of the targeted competencies. These findings provide preliminary support for the potential interchangeability of AI-generated items when aligned with clearly defined learning outcomes. However, two exceptions merit closer examination. First, the item pair corresponding to LO6 (*Lists possible sources of random error based on a given scenario*) showed no significant correlation. Both items, Item 14 (ChatGPT) and Item 3 (DeepSeek), were also characterized by low discrimination indices, suggesting that the items may lack sufficient capacity to effectively differentiate between higher- and lower-performing examinees. This may explain the lack of association and highlights the importance of item quality in establishing inter-item consistency. Second, a weak but statistically significant correlation was found between the items representing LO8 (*Selects the appropriate reliability estimation technique based on a given scenario*). Similar to LO6, one of the items (Item 19) exhibited suboptimal discrimination, which may have contributed to the attenuated correlation. In contrast, item pairs associated with LO3 (*Differentiates among nominal, ordinal, interval, and ratio scales based on a given scenario*) and LO9 (*Compares different types of tests in terms of guessing probability*) exhibited statistically significant correlations that approached a moderate level. These findings suggest that, overall, both AI tools demonstrated a consistent capacity to generate items aligned with the same cognitive intent, particularly when the psychometric quality of the items (especially their discrimination indices) was adequate.

The low correlations observed for certain item pairs may be attributed to factors such as the cognitive complexity or abstract nature of the targeted LOs. These characteristics might lead AI tools to interpret prompts differently, resulting in items that tap into distinct facets of the same construct. For instance, LO6 (*Lists possible sources of random error based on a given scenario*) or LO8 (*Selects the appropriate reliability estimation technique based on a given scenario*) involve subtle conceptual distinctions that may not be uniformly addressed by different AI models. Additionally, variation in how each tool formulates distractors and key terms might have influenced how students interpreted and responded to the items, reducing response consistency. These inconsistencies suggest that, for certain types of outcomes, AI-generated items may require closer scrutiny and expert validation to ensure alignment and comparability. This suggests that AI tools may interpret the same prompt differently and highlights the importance of careful prompt engineering and post-hoc expert review.

At the test level, reliability analyses indicated that the 20-item trial test demonstrated an almost acceptable level of internal consistency for a classroom-based assessment, with a KR-20 coefficient of .69 and a Spearman-Brown corrected split-half reliability of .73 (Rudner & Schafer, 2002). Three items fell below the acceptable discrimination threshold of .30: two from ChatGPT (Items 14 and 19) and one from DeepSeek (Item 3). Although removing these items did not significantly affect the KR-20 coefficient, it enhanced split-half reliability. When the

final 17-item version was analyzed, the split-half reliability increased to .75. This result suggests that the refined test version demonstrated an acceptable level of internal consistency for a teacher-made assessment, particularly in the context of classroom-based evaluation (Murphy & Davidshofer, 1991). Additionally, a comparison of student average scores across the two test halves revealed no statistically significant difference, indicating that the final form exhibited balanced performance and item difficulty across both halves.

The correlation analysis revealed that both the ChatGPT and DeepSeek generated item halves were strongly associated with the total test score, indicating meaningful contributions from both AI tools to the overall construct being measured. However, a statistically significant difference between the two correlations was observed, with the DeepSeek half showing a higher correlation with the total score. Given the absence of significant psychometric differences at the item level, this marginally stronger association may not reflect a substantive difference in item quality but rather a minor variation in how well the items collectively align with the latent trait assessed by the full test. One possible explanation for this finding is that the slight advantage in performance may not stem from psychometric deficiencies in the ChatGPT items, but rather from subtle nuances in the generative processes of the two AI models. Specifically, the variation may emerge from differences in how each model interprets prompts, formulates item stems and distractors, or aligns content with LOs. These small variations, although not reflected in item-level statistics, can accumulate and manifest in stronger associations with total test performance.

Since expert judgment plays a fundamental role in establishing content validity during the test development process (Crocker & Algina, 1986), expert reviews served to confirm the alignment between the items and their corresponding LOs and to identify items that may require modification. The fact that only four out of twenty items were recommended for revision, with suggestions evenly distributed across ChatGPT and DeepSeek, indicates that both AI tools were able to generate items that largely met expert expectations. An analysis of the expert feedback revealed that one suggestion focused on restructuring the correct answer, another on revising the item stem, while the remaining recommendations addressed the distractor options. This pattern indicates that the majority of expert critiques were concerned with the plausibility, clarity, or functionality of the distractors rather than fundamental flaws in item construction or alignment with the LOs. Such feedback suggests that AI tools are generally capable of generating items with valid stems and correct answers, but may still require human refinement in designing high-quality distractors. Distractors are designed to reduce the likelihood that students with insufficient knowledge of the targeted construct will correctly answer the item, thereby making it more difficult for those who do not know the answer to earn credit from the item (Özçelik, 2013). Accordingly, well-constructed distractors contribute to the item's ability to differentiate between higher- and lower-performing examinees. These findings reinforce the role of expert review in augmenting the psychometric quality of AI-generated items, particularly by enhancing the performance of distractors that may otherwise reduce item difficulty or fail to capture student misconceptions.

The results of the G study indicated that the experts evaluated the items in a consistent manner, suggesting an agreement among raters. For the evaluations provided by three experts on the 20 items generated by the AI tools both the *G* and *Phi* coefficients were calculated to be .64. According to Shavelson and Webb (1991), *G* and *Phi* coefficients must be at least .80 to be considered “high.” However, Brennan (2004) noted that the adequacy criteria for interpreting these coefficients may vary depending on the context (as cited in Atılğan, 2004). In this regard, although the coefficients obtained from three field experts do not reach the threshold for being considered high, they nonetheless indicate a certain level of agreement. The coefficients estimated through various scenarios with different numbers of experts increased with the number of experts. Based on these results, it was concluded that a minimum of seven expert raters are required to achieve optimal reliability in evaluating AI-generated items.



This study contributes to the growing literature on AI-assisted test development by empirically evaluating the psychometric properties of multiple-choice items generated by ChatGPT and DeepSeek. The findings indicate that both tools are capable of producing test items with acceptable levels of difficulty, discrimination, and reliability. Even though DeepSeek items appeared to perform slightly better in terms of item-level psychometric qualities, this difference was not statistically significant, implying that both tools may serve equally well in classroom contexts. However, the study also highlights the need for expert involvement, especially in ensuring content validity and alignment with LOs. The findings obtained through the G and D studies provided valuable insight into the reliability of expert evaluations of AI-generated test items. While the initial reliability coefficients calculated based on three raters ( $G = .64$ ;  $\Phi = .64$ ) indicated moderate consistency, the scenario-based simulations revealed that reliability improves substantially as the number of raters increases. Specifically, the results showed that to ensure a high level of reliability in the expert evaluation process, a minimum of seven experts is required. This highlights the importance of involving a sufficient number of qualified raters when validating content in AI-assisted test development processes.

## 5. LIMITATIONS and FUTURE RESEARCH

While the findings of this study provide valuable insights into the psychometric potential of AI-generated multiple-choice items, several limitations should be acknowledged. First, the study was limited to AI-based automatic item generation using natural language processing, rather than template-based AIG. This decision was made to avoid structural constraints in item generation and to offer a more accessible approach for teachers and practitioners. While AI-based AIG emphasizes the use of large language models to generate items from open-ended prompts, template-based AIG allows content experts to construct items by manipulating specific components within a structured framework. Although more flexible, the AI-based approach often produces technically-oriented items that require further refinement and expert validation before use in educational assessment contexts (Graesser *et al.*, 2012). Second, the LOs in this study were limited to the Comprehension level of Bloom's taxonomy in order to ensure that the measured constructs corresponded to LOs that can be effectively assessed through multiple-choice items and to avoid introducing an additional source of variability that might result from using LOs across different taxonomic levels. Future research could explore whether AI tools are equally capable of generating high-quality items across different cognitive levels, such as knowledge and application. Another limitation of this study is that no distractor analysis was conducted. Future research is recommended to conduct an in-depth examination of AI tools' performance in generating plausible and effective response options.

Additionally, the study was confined to a single course in the field of educational measurement. The generalizability of the findings to other disciplines remains an open question and warrants further investigation. Third, the study employed a relatively small item pool ( $n = 20$ ) and a single round of expert review. Expanding the number of items and incorporating iterative review cycles may yield more robust findings. Since this study focused on the number of experts, in future research, expert background variables (such as level of experience or content area) may also be included as a source of variance in the G study. Additionally, future research may manipulate the number of items in the D study to estimate Phi and G coefficients under different scenarios. Future studies could involve multiple student cohorts across different institutions to assess the stability of psychometric results over time and context. Finally, although two prominent AI tools were compared, the field of large language models is rapidly evolving. Ongoing comparisons involving newer and more diverse models will be essential to understanding the full potential and limitations of AI in educational assessment.

Based on these findings, the following recommendations can be made: AI tools can be used effectively for initial item generation, but expert review is crucial to ensure validity and clarity. Future studies should explore AI performance across different cognitive levels of Bloom's

Taxonomy and in other subject areas. Training educators in prompt engineering and item validation may enhance the practical use of AI in assessment development. In conclusion, AI-based item generation holds significant potential for supporting teachers in creating reliable and valid assessments. However, it should be seen as a complementary rather than a replacement tool; human expertise remains essential in ensuring the educational quality of test items. Although D study simulations helped identify the optimal number of experts, future studies should replicate this analysis with a larger and more diverse panel of experts, as the current study was limited to a relatively small and homogeneous sample of experts. Additionally, the item pool was restricted to a specific course context (Educational Measurement and Evaluation), and the findings may not be generalizable to other subject areas or educational levels. Future studies should investigate the reliability and validity of AI-generated test items across different disciplines, cognitive levels, and learner populations. Moreover, qualitative analyses of expert feedback could provide deeper insights into the nature of item flaws and the limitations of AI tools in interpreting instructional prompts. Expanding the scope and scale of similar studies will contribute to a more comprehensive understanding of the role of artificial intelligence in educational assessment design.

### Acknowledgments

No funds, grants, or other support were received for this study. Moreover, the study has not been previously presented at a conference or a scholarly meeting.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number:** Aksaray University, 21/01/2025-03.

### Contribution of Authors

**Ceylan Gündeğer Kılıcı** was solely responsible for the conception and design of the study, development of research materials, data collection, statistical analysis, interpretation of results, and the preparation of the manuscript.

### Orcid

Ceylan Gündeğer Kılıcı  <https://orcid.org/0000-0003-3572-1708>

### REFERENCES

- Alafnan, M.A. (2025). DeepSeek vs. ChatGPT: A comparative evaluation of AI tools in composition, business writing, and communication tasks. *Journal of Artificial Intelligence and Technology*, 5, 202-210. <https://doi.org/10.37965/jait.2025.0740>
- Anderson, L.W., & Krathwohl, D.R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. Longman.
- Atılğan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma [A research on comparisons of generalizability theory and many facets Rasch measurement]* [Unpublished doctoral dissertation]. Hacettepe University.
- Atılğan, H. (2009). Madde ve test istatistikleri [Item and test statistics]. In H. Atılğan, A. Kan, & N. Doğan (Eds.), *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (3<sup>rd</sup> ed., pp. 293-314). Anı Publishing.
- Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması [Measurement in education and psychology: Classical test theory and its applications]*. ÖSYM Publications.
- Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans, Green.

- Brennan, R.L. (2001). *Generalizability Theory*. Springer-Verlag.
- Clauser, B. (2008). A review of the EDUG software for generalizability analysis. *International Journal of Testing*, 8(3), 296-301. <https://doi.org/10.1080/15305050802262357>
- Clauser, B.E., Margolis, M.J., & Case, S.M. (2006). Testing for licensure and certification in the professions. In R. L. Brennan (Ed.), *Educational measurement* (4th ed. pp. 701-730). Praeger Publications.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston Inc.
- Daily Sabah. (2023, October 06). *Türkiye's student assessment center to use AI to set exam question*. <https://www.dailysabah.com/turkiye/education/turkiyes-student-assessment-center-to-use-ai-to-set-exam-question>
- Dancey, C.P., & Reidy, J. (2017). *Statistics without maths for psychology*. Pearson.
- DeepSeek AI. (2024). *DeepSeek-v2: Advancing open-source large language models*. <https://www.deepseek.com>
- Dienhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, 10(4), Article e0121945. <http://dx.doi.org/10.1371/journal.pone.0121945>
- Downing, S.M., & Haladyna, T.M. (2011). *Handbook of test development*. Lawrence Erlbaum Associates Publishers.
- Educational Testing Service. (2025a). *e-rater® scoring engine*. ETS. <https://www.ets.org/erater.html>
- Educational Testing Service. (2025b). *SpeechRater® scoring engine*. ETS. <https://www.ets.org/speechrater.html>
- Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education*. McGraw-Hill.
- Gao, H., Hashim, H., & Md Yunus, M. (2025). Assessing the reliability and relevance of DeepSeek in EFL writing evaluation: A generalizability theory approach. *Language Testing in Asia*, 15(33). <https://doi.org/10.1186/s40468-025-00369-6>
- Gierl, M.J., & Haladyna, T.M. (2013). *Automatic item generation: Theory & practice*. Routledge.
- Gierl, M.J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge.
- Gierl, M.J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education*, 46(8), 757-765. <https://doi.org/10.1111/j.1365-2923.2012.04289.x>
- Graduate Management Admission Council. (2009). *Fairness of automated essay scoring of GMAT AWA*. <https://www.gmac.com/market-intelligence-and-research/research-library/gmat-test-taker-data/research-reports-gmat-related/fairness-of-automated-essay-scoring-of-gmat-awa>
- Graesser, A.C., Conley, M.W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, vol. 3. application to learning and teaching* (pp. 451-473). American Psychological Association. <https://doi.org/10.1037/13275-018>
- Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. Routledge.
- Irwing, P., & Hughes, D.J. (2018). Test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development (First ed., pp. 3-48)*, John Wiley & Sons Ltd. <https://doi.org/10.1002/9781118489772.ch1>

- Kanık, M. (2024). The use of ChatGPT in assessment. *International Journal of Assessment Tools in Education*, 11(3), 608-621. <https://doi.org/10.21449/ijate.1379647>
- Kendall, M.G., & Smith, B.B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3), 275-287. <http://www.jstor.org/stable/2235668>
- Kıyak, Y.S., & Emekli, E. (2024). ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: A literature review. *Postgraduate Medical Journal*, 100(1189), 858-865. <https://doi.org/10.1093/postmj/qgae065>
- Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174. <https://doi.org/10.2307/2529310>
- Leslie, T., & Gierl, M.J. (2023). Using automatic item generation to create multiple-choice questions for pharmacy assessment. *American Journal of Pharmaceutical Education*, 87(10), 1-7. <https://doi.org/10.1016/j.ajpe.2023.100081>
- Liu, Z., He, X., Liu, L., Liu, T., & Zhai, X. (2023). Context matters: A strategy to pre-train language model for science education. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O.C. Santos (Eds.), *Artificial intelligence in education*. Springer. [https://doi.org/10.1007/978-3-031-36336-8\\_103](https://doi.org/10.1007/978-3-031-36336-8_103)
- Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), Article 410. <https://doi.org/10.3390/educsci13040410>
- Makowski, D., Wiernik, B., Patil, I., Lüdecke, D., & Ben-Shachar, M. (2022). *correlation: Methods for Correlation Analysis* (Version 0.8.3) [R package]. <https://CRAN.R-project.org/package=correlation>
- Malik, M., Rehan, S., Zimbittas, G., & Manna, S. (2024). Multiple-choice questions reimaged: Exploring the ethical and pedagogical implications of GenAI for higher education. In T. Fujita (Ed.), *Proceedings of the British Society for Research into Learning Mathematics (BSRLM)*, 44(1). <https://bsrlm.org.uk/wp-content/uploads/2024/05/BSRLM-CP-44-1-05.pdf>
- Meng, X.-L., Rosenthal, R., & Rubin, D.B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, 111(1), 172-175. <https://doi.org/10.1037/0033-2909.111.1.172>
- Murphy, K.R., & Davidshofer, C.O. (1991). *Psychological testing: Principles and applications*. Prantice Hall.
- Ngo, A., Gupta, S., Perrine, O., Reddy, R., Ershadi, S., & Remick, D. (2024). ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Academic Pathology*, 11(1), 1-5. <https://doi.org/10.1016/j.acpath.2023.100099>
- OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. [https://www.oecd.org/en/publications/pisa-2022-results-volume-i\\_53f23881-en/full-report/adaptive-testing-in-pisa-2022\\_21364c8d.html](https://www.oecd.org/en/publications/pisa-2022-results-volume-i_53f23881-en/full-report/adaptive-testing-in-pisa-2022_21364c8d.html)
- OpenAI. (2023). *GPT-4 technical report*. <https://openai.com/research/gpt-4>
- ÖSYM (2024, February 13). *ÖSYM Başkanı Ersoy: Yapay zekâ ile soru üreteceğiz* [ÖSYM President Ersoy: We will generate questions with artificial intelligence]. <https://www.osym.gov.tr/TR,29174/osym-baskani-ersoy-yapay-zeka-ile-soru-uretecegiz-13022024.html>
- Özçelik, D.A. (1992). *Ölçme ve değerlendirme [Measurement and assessment]*. ÖSYM Publication.
- Özçelik, D.A. (2013). *Test hazırlama kılavuzu [Test preparation manual]*. Pegem.
- R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2). R Foundation for Statistical Computing. <https://www.R-project.org/>
- Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.3.9) [R package]. Northwestern University. <https://CRAN.R-project.org/package=psych>



- Robitzsch, A. (2024). *sirt: Supplementary Item Response Theory Models* (Version 4.1.-15). [R package]. <https://CRAN.R-project.org/package=sirt>
- Rudner, L., & Schafer, W. (2002) *What teachers need to know about assessment*. National Education Association.
- Rycroft-Smith, L., & Macey, D. (2024). Using AI for question generation in mathematics education: What are the advantages and disadvantages? In T. Fujita (Ed.), *Proceedings of the British Society for Research into Learning Mathematics (BSRLM)*, 44(1). <https://bsrlm.org.uk/wp-content/uploads/2024/05/BSRLM-CP-44-1-07.pdf>
- The Princeton Review. (2024). *Digital SAT security and fairness*. <https://www.princetonreview.com/college-advice/digital-security-and-fairness>
- Schober, P., Boer, C., & Schwarte, L.A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5), 1763-1768.
- Ratner, B. (2009) The correlation coefficient: Its values range between +1/-1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*, 17, 139-142. <https://doi.org/10.1057/jt.2009.5>
- Seldon, A., & Abidoye, O. (2018). *The fourth education revolution*. University of Buckingham Press.
- Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. SAGE Publications.
- Shin, D. (2023). A case study on English test item development training for secondary school teachers using AI tools: Focusing on ChatGPT. *Language Research*, 59(1), 21-42. <https://doi.org/10.30961/lr.2023.59.1.21>
- Society, S., & Group, E.W. (2010). *Edug user guide*. Edumetrics.
- Tan, Ş. (2014). *Öğretimde ölçme ve değerlendirme: KPSS el kitabı [Assessment and evaluation in instruction: KPSS handbook]*. Pegem.
- Thorndike, R.M., & Thorndike-Christ, T. (2014). *Measurement and evaluation in psychology and education*. Pearson.
- Turgut, M.F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Pegem.
- Urhan, S., Gençaslan, O., & Dost, Ş. (2024). An argumentation experience regarding concepts of calculus with ChatGPT. *Interactive Learning Environments*, 32(10), 7186-7211. <https://doi.org/10.1080/10494820.2024.2308093>
- Ünal, D., Erdem, Z.Ç., & Şahin, Z.G. (2025). Will artificial intelligence succeed in passing this test? Creating an achievement test utilizing ChatGPT. *Education & Information Technologies*, 30, 17263-17287. <https://doi.org/10.1007/s10639-025-13461-4>
- Xiaoyu, W. (2024, June 21). AI Scores High in Gaokao Language Tests, Low in Math. China Daily. <https://www.chinadaily.com.cn/a/202406/21/WS6674bb00a31095c51c50a0a9.html>
- Wang, J., & Heung, K. (2025). Educational innovation driven by artificial intelligence: The impact of DeepSeek on teachers' teaching models. *Learning & Education*, 14(1), 38-42. <https://ojs.piscomed.com/index.php/L-E/article/view/4291>
- Wardat, Y., Tashtoush, M.A., Alali, R., & Jarrah, A.M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), Article em2286. <https://doi.org/10.29333/ejmste/13272>
- Wickham, H., & Bryan, J. (2023). readxl: Read Excel files (Version 1.4.3) [R package]. <https://CRAN.R-project.org/package=readxl>
- Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.4) [R package]. <https://CRAN.R-project.org/package=dplyr>
- Willse, J.T. (2014). *CTT: Classical Test Theory functions* (Version 2.3.3) [R package]. <https://CRAN.R-project.org/package=CTT>



- Yogesh, A., Telon, G., Lovely, T.F., & Braiton, M. (2025). A comparative study: Evaluating ChatGPT and DeepSeek AI tools in practice. *International Journal of Open Information Technologies*, 13(5), 67-70. <https://cyberleninka.ru/article/n/a-comparative-study-evaluating-chatgpt-and-deepseek-ai-tools-in-practice>
- Zhai, X. (2025). DeepSeek: Transforming the Foundations of Education. Preprints. <https://doi.org/10.20944/preprints202503.1776.v1>

## APPENDICES

### Appendix 1. Prompt.

Could you generate an appropriate multiple-choice test item for each of the learning outcomes I will provide?

- Faculty: Faculty of Education
- Course: Educational Measurement and Evaluation
- Level: Undergraduate
- The questions should have only one correct answer and five answer options in total.

Learning outcome 1: Distinguishes between quantitative, qualitative, continuous, discrete, dependent, and independent variables based on a given scenario. (Comprehension level)

Learning outcome 2: Identifies direct, indirect, and derived types of measurement based on a given scenario. (Comprehension level)

Learning outcome 3: Differentiates among nominal, ordinal, interval, and ratio scales based on a given scenario. (Comprehension level)

Learning outcome 4: Breaks down the concept of evaluation into its components: measurement, criterion, and decision, by using a given scenario. (Comprehension level)

Learning outcome 5: Distinguishes between constant, systematic, and random error types based on a given scenario. (Comprehension level)

Learning outcome 6: Lists possible sources of random error based on a given scenario. (Comprehension level)

Learning outcome 7: Interprets a correlation matrix involving five different variables. (Comprehension level)

Learning outcome 8: Selects the appropriate reliability estimation technique based on a given scenario. (Comprehension level)

Learning outcome 9: Compares different types of tests in terms of guessing probability. (Comprehension level)

Learning outcome 10: Distinguishes among the three different meanings of reliability—stability, consistency, and sensitivity—based on a given scenario. (Comprehension level)

**Appendix 2.** *Tetrachoric correlations coefficients between the test items and p-values.*

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20
Item 1	1.00																			
p	.00																			
Item 2	.52	1.00																		
p	.00	.00																		
Item 3	.05	.06	1.00																	
p	.58	.54	.00																	
Item 4	.10	.30	.17	1.00																
p	.28	.00	.07	.00																
Item 5	.29	.22	.29	.45	1.00															
p	.00	.01	.00	.00	.00															
Item 6	.27	.36	-.24	.21	.47	1.00														
p	.00	.00	.01	.03	.00	.00														
Item 7	.34	.26	.36	.10	.25	-.09	1.00													
p	.00	.01	.00	.27	.01	.34	.00													
Item 8	.16	.30	.13	.58	.18	-.12	.12	1.00												
p	.09	.00	.15	.00	.05	.19	.20	.00												
Item 9	.21	.21	-.02	.31	.25	.05	.12	.25	1.00											
p	.02	.02	.85	.00	.01	.59	.20	.01	.00											
Item 10	.35	.27	.20	.31	.38	.00	.39	.28	.17	1.00										
p	.00	.00	.03	.00	.00	.96	.00	.00	.07	.00										
Item 11	.19	.33	.25	.07	.01	.08	.22	.07	-.01	.49	1.00									
p	.04	.00	.01	.45	.96	.41	.02	.47	.92	.00	.00									
Item 12	.05	.31	.58	.23	.35	.06	.38	.03	.38	.42	.38	1.00								
p	.62	.00	.00	.01	.00	.48	.00	.71	.00	.00	.00	.00								
Item 13	.10	.16	.13	.47	.12	-.05	.23	.09	.30	.19	.36	.13	1.00							
p	.29	.07	.17	.00	.21	.60	.01	.30	.00	.04	.00	.16	.00							
Item 14	-.32	.12	-.09	.42	.16	.02	.04	.21	-.25	.25	-.05	-.06	.22	1.00						
p	.00	.19	.31	.00	.08	.83	.65	.02	.01	.01	.56	.55	.01	.00						
Item 15	.45	.17	.44	.26	.18	-.09	.18	.18	.20	.27	.35	.27	.21	-.36	1.00					
p	.00	.06	.00	.00	.05	.33	.05	.05	.03	.00	.00	.00	.02	.00	.00					
Item 16	.30	-.04	.13	.54	.12	.30	-.15	.27	.25	.15	.12	-.01	.27	-.05	.27	1.00				
p	.00	.69	.14	.00	.18	.00	.10	.00	.01	.09	.20	.89	.00	.57	.00	.00				
Item 17	.30	.33	.24	.16	.17	.30	-.00	-.08	.08	.15	-.04	.54	.00	-.18	.31	.11	1.00			
p	.00	.00	.01	.09	.06	.00	.97	.41	.36	.11	.66	.00	.96	.05	.00	.23	.00			
Item 18	.23	.04	.10	.27	.13	-.01	-.03	.07	-.03	.39	.16	.20	.26	.16	.35	.31	.23	1.00		
p	.01	.64	.27	.00	.17	.93	.71	.48	.72	.00	.08	.03	.00	.09	.00	.00	.01	.00		
Item 19	.43	.38	-.29	.03	-.10	.10	.04	.18	.13	.10	.10	-.11	-.09	-.13	.10	.19	.19	-.16	1.00	
p	.00	.00	.00	.71	.28	.27	.69	.04	.15	.27	.27	.23	.35	.15	.26	.04	.03	.08	.00	
Item 20	.31	.16	.25	.17	.26	.35	.16	.04	.21	.13	.37	.43	.15	.02	.41	.42	.23	.37	.09	1.00
p	.00	.09	.01	.06	.00	.00	.07	.63	.02	.16	.00	.00	.10	.79	.00	.00	.01	.00	.35	.00

**Appendix 3.** *KR-20 if item deleted.*

	Scale Mean	Scale Variance	KR-20
Item 1	13.91	9.008	.667
Item 2	14.05	8.787	.665
Item 3	13.78	9.684	.685
Item 4	13.87	9.024	.665
Item 5	14.12	8.860	.670
Item 6	14.21	9.292	.688
Item 7	13.83	9.574	.685
Item 8	14.08	9.237	.685
Item 9	13.87	9.461	.683
Item 10	13.84	9.176	.670
Item 11	13.88	9.354	.680
Item 12	13.78	9.448	.676
Item 13	13.87	9.427	.682
Item 14	13.79	9.948	.696
Item 15	13.94	9.030	.670
Item 16	13.78	9.549	.680
Item 17	14.28	9.159	.682
Item 18	14.12	9.163	.682
Item 19	13.90	9.654	.693
Item 20	14.10	8.763	.665
Total	14.68	10.110	.690

**Appendix 4.** *Expert evaluation for the items generated by ChatGPT and DeepSeek.*

Item No	AI	Expert 1	Expert 2	Expert 3	Revision
1	ChatGPT	3	3	3	No
2	DeepSeek	3	3	3	No
3	DeepSeek	3	3	3	No
4	ChatGPT	3	3	3	No
5	DeepSeek	3	3	3	No
6	ChatGPT	3	3	3	No
7	ChatGPT	3	3	3	No
8	DeepSeek	3	3	3	No
9	DeepSeek	3	3	3	No
10	ChatGPT	3	3	3	No
11	DeepSeek	3	3	3	No
12	ChatGPT	3	3	3	No
13	DeepSeek	3	3	3	No
14	ChatGPT	2	2	3	Yes
15	DeepSeek	2	3	2	Yes
16	ChatGPT	3	3	3	No
17	DeepSeek	2	3	2	Yes
18	ChatGPT	3	3	3	No
19	ChatGPT	3	3	2	Yes
20	DeepSeek	3	3	3	No

Kendall's  $W = .58$



## Construction and validation of a multilingual diagnostic instrument for neuromyths and their origins

Oktay Cem Adigüzel<sup>1,2\*</sup>, Sibel Küçükkayhan<sup>2</sup>, Patrice Potvin<sup>3</sup>, Derya Atik-Kara<sup>2</sup>

<sup>1</sup>Université du Québec à Trois-Rivières, Faculty of Education (UQTR), Department of Educational Sciences, Canada

<sup>2</sup>Anadolu University, Faculty of Education, Department of Educational Sciences, Türkiye

<sup>3</sup>Université du Québec à Montréal (UQAM), Faculty of Education, Department of Didactic, Canada

### ARTICLE HISTORY

Received: Oct. 12, 2024

Accepted: Sep. 3, 2025

### Keywords:

Neuromyths,  
Learning and teaching,  
Neuroscience,  
Teacher training,  
Assessment tool.

**Abstract:** This study presents the development of a comprehensive neuromyth identification tool designed to be valid, reliable, and multilingual, including French, English, Turkish, Greek, Kazakh, Arabic, Malay, and Chinese. By incorporating languages from diverse geographic regions, the tool aims to increase the accessibility and relevance of neuromyth research, allowing for more comprehensive and generalizable findings. The primary research question guiding this study was: "What structural properties should a valid and reliable instrument have to effectively identify teachers' primary neuromyth beliefs and the origins of these beliefs?" A mixed-methods approach was used, integrating both quantitative and qualitative methods to ensure the robustness of the instrument. The development process unfolded in four key stages: (1) a thorough literature review to identify existing neuromyths and relevant survey instruments, (2) the design of the initial questionnaire, (3) pilot testing to evaluate and refine the instrument, and (4) language adaptation to ensure cultural and linguistic appropriateness in the target languages. The resulting neuromyth identification tool has been rigorously tested for its structural properties, such as validity and reliability, across different linguistic and cultural contexts.

## 1. INTRODUCTION

The brain's intricate and often mystifying properties have long made it a subject of fascination and commercial exploitation (Dekker *et al.*, 2012; Ferrero *et al.*, 2016). Beliefs about how the brain functions that are commonly held but not based on facts are called neuromyths. These neuromyths can arise from misunderstandings, misinterpretations, or oversimplifications of neuroscientific findings (OECD, 2002). These misconceptions have been spread through informal sources such as films, documentaries, advertisements, and social media platforms (Carter *et al.*, 2020; Howard-Jones, 2014; Ruhaak & Cook, 2018). However, studies have shown that formal education systems, especially teacher training programs, professional

\*CONTACT: Oktay Cem ADIGÜZEL ✉ [ocadiguzel@anadolu.edu.tr](mailto:ocadiguzel@anadolu.edu.tr) 📧 Université du Québec à Trois-Rivières, Faculty of Education (UQTR), Department of Educational Sciences, Canada

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

development activities, and school curricula, perpetuate these myths by incorporating them into teaching methods (Tardif *et al.*, 2015; Vig *et al.*, 2023; Zhang *et al.*, 2019).

As defined by the Organisation for Economic Co-operation and Development (OECD, 2002), neuromyths stem from the misunderstanding, misinterpretation, or misapplication of neuroscientific findings regarding the cognitive and affective characteristics of learning. A common example is the concept of learning styles, which posits that individuals have distinct preferences, such as visual, auditory, or kinesthetic learning. Although this idea is based on the observation that different types of information are processed in specialized regions of the brain, it has not been scientifically validated by neuroscience research (Tardif *et al.*, 2015; Zhang *et al.*, 2019). Such misconceptions are reinforced through popular media, commercial “brain-based” programs, and even formal teacher education curricula, thereby becoming entrenched in educational settings (Vig *et al.*, 2023; Zhang *et al.*, 2019). Consequently, without a critical evaluation of these claims against neuroscientific evidence, teachers may inadvertently integrate them into their instructional practices.

In addition to learning styles, other widely held neuromyths include the theory of multiple intelligences, the notion of left- and right-brain dominance, the belief in effective multitasking, and the claim that humans use only 10% of their brains. The persistence of these myths highlights the importance of accurately interpreting scientific evidence and recognizing the brain’s complex, holistic role in learning. The prevalence of these neuromyths in education has significantly influenced our understanding of learning, often with negative consequences. Misconceptions about brain training, learning styles, nutrition, and medication have led to inappropriate practices that threaten learning, achievement and overall well-being of students. Despite evidence to the contrary, such practices appear to persist largely in educational settings (Ching *et al.*, 2020). Research has shown that socio-cognitive biases are strong predictors of neuromyth beliefs, suggesting that modifying thought schema is essential to dispel these myths (van Elk, 2019).

Neuromyths spread rapidly through both formal and informal educational environments, facilitated by the misinterpretation of neuroscience research, exploitation by those unfamiliar with the field, and the appeal of sensationalist narratives in popular media. Teachers play a critical role in both perpetuating and dispelling neuromyths (Dekker *et al.*, 2012). During both pre-service and in-service training, teachers engage in formal and informal learning activities to support their professional development. In this process, certain practices that emerge as innovative approaches are often adopted without undergoing rigorous scientific scrutiny, inadvertently leading to the creation of new neuromyths or the reinforcement of existing ones (Ching *et al.*, 2020; Dekker *et al.*, 2012; Tardif *et al.*, 2015; Zhang *et al.*, 2019). The study conducted by Ching *et al.* (2020) highlights the central role of educators in perpetuating neuromyths. Therefore, it is essential that teachers evaluate information from various disciplines with constructive skepticism and actively seek scientific validation.

Although studies in literature identify common neuromyths, they offer minimal insight into their origins and dissemination. Additionally, research indicates that the data collection instruments used to identify neuromyths have certain scientific methodology weaknesses. These weaknesses include insufficient reporting of the instruments’ structural characteristics, inadequate evidence of adaptation to linguistic and cultural contexts, and uncertainties in sample selection. Neuromyths significantly impact the learning process. Therefore, it is important to determine the neuromyth beliefs and learning sources of educational stakeholders, particularly teachers and pre-service teachers, regarding advances in neuroscience. This can be achieved by using well-designed instruments developed in accordance with rigorous scientific methods.

In recent years, researchers have conducted multiple studies to determine the prevalence of neuromyths among teachers. These research projects have employed various measurement

instruments. A review of this research reveals that data are most often collected using true/false or Likert-type questionnaires. One widely used instrument is the neuromyth questionnaire developed by Dekker *et al.* (2012), which has been adapted and used in many countries. It typically consists of items related to neuromyths and mental functions with true/false or "don't know" response options and contains between 12 and 35 items. The tool used in the study of Dekker *et al.* (2012) is based on the OECD's 2002 report on neuromyths and the myths identified therein.

Researchers have generally used these measurement instruments in translation without linguistic and cultural adaptation. However, only a few studies have adapted them for specific contexts. For instance, Schmitt *et al.* (2023) expanded the study of Dekker *et al.* (2012) and Macdonald *et al.* (2017) based on a 23-item scale by adding nine new giftedness-related items. They presented the questionnaire in English, French, and German to accommodate Luxembourg's multilingual structure and tested the translations for cultural validity. Another distinctive approach among these studies is the scenario-based instrument developed by Tovazzi *et al.* (2020) titled the Neuroscience Against Neuromyths Questionnaire (NNQ). Unlike traditional neuromyth questionnaires, this tool provides a more practical and contextualized approach, evaluating how teachers would respond to real classroom scenarios.

Structurally, it is observed that the reliability of the items in most of these instruments has been evaluated using Cronbach's alpha coefficients. Although valid and reliable data collection instruments are essential for accurately analyzing research data, the literature lacks validity and reliability analyses during the development of these instruments. Consequently, there is insufficient information on item analysis, pilot testing, and other development processes. [Table 1](#) presents a sample based on several neuromyth studies conducted in recent years, with the most recent study serving as the starting point.

As shown in [Table 1](#), most studies have used instruments derived from one or two primary sources. Linguistic adaptations to target languages are rarely undertaken, and the structural characteristics of the instruments are rarely verified or reported. Overall, the tools used to assess neuromyths exhibit significant methodological variation. While most rely on the traditional true–false test format, some studies have incorporated scenario-based items, multiple-choice questions, or Likert-type scales to enrich their assessments. Within the literature, instruments explicitly designed to gather data on the origins of neuromyths remain particularly scarce. The overwhelming focus of existing research has been on the identification of neuromyths themselves, with insufficient attention devoted to systematically investigating their sources.

An effective diagnosis of neuromyths and their origins requires a precise conceptualization of the myths, a scientific plan for linguistic and cultural adaptation processes, and a thorough report on the validity and reliability of the instruments. Such methodological rigor would facilitate more accurate identification and support the development of targeted intervention programs, thereby enhancing the capacity to distinguish between scientific evidence and popular misconceptions within educational contexts. To address these gaps, this study aims to create a neuromyth identification tool that is valid, reliable, and multilingual encompassing French, English, Turkish, Greek, Kazakh, Arabic, Malay, and Chinese languages. By including languages spoken across a broad geographic area, this tool is expected to enhance the usability and applicability of the surveys, facilitating more comprehensive and generalizable research findings.

Thus, the research question guiding this study is: "What structural properties could a valid and reliable instrument possess to identify teachers' primary neuromyth beliefs and the origins of these beliefs?" As previously discussed, the lack of such comprehensive instruments in the existing literature emphasizes the novelty and uniqueness of this research. Therefore, the objective of this study is to conduct the development of a questionnaire that facilitates critical reflection on the aforementioned properties.

**Table 1.** Comparative review of neuromyth identification tools.

Source	Instrument Description	Adapted / Based On	Language/Cultural Adaptation	Validity / Reliability Reported
Sazaka <i>et al.</i> , 2024	5-item neuromyth identification tool	Tardif <i>et al.</i> (2015)	Not reported	Not reported
Deibl <i>et al.</i> , 2023	46-item neuromyth identification tool	Dekker <i>et al.</i> (2012); Krammer <i>et al.</i> (2019, 2020)	Not reported	Not reported
Vig <i>et al.</i> , 2023	23-item neuromyth identification tool	Dekker <i>et al.</i> (2012)	Not reported	Not reported
Deans <i>et al.</i> , 2022	5-item neuromyth identification tool	Dekker <i>et al.</i> (2012)	Not reported	Not reported
Jeyavel <i>et al.</i> , 2022	7-item neuromyth identification tool	Not reported	Not reported	Not reported
Ruiz-Martin <i>et al.</i> , 2022	32-item neuromyth identification tool	Dekker <i>et al.</i> (2012)	Not reported	Not reported
Simoes <i>et al.</i> , 2022	30-item neuromyth identification tool	Not reported	Not reported	Not reported
Bisessar, 2021	30-item neuromyth identification tool	OECD (2002)	Not reported	Not reported
Craig <i>et al.</i> , 2021	30-item neuromyth identification tool	Dekker <i>et al.</i> (2012)	Not reported	Not reported
Carter <i>et al.</i> , 2020	32-item neuromyth identification tool	Dekker <i>et al.</i> (2012)	Not reported	Not reported
Ching <i>et al.</i> , 2020	17-item neuromyth identification tool	Dekker <i>et al.</i> (2012),	Translation & back-translation	Not reported
McMahon <i>et al.</i> , 2019	32-item neuromyth belief questionnaire	Dekker <i>et al.</i> (2012)	Not reported	Not reported
Zhang <i>et al.</i> , 2019	40-item neuromyth identification tool	Dekker <i>et al.</i> (2012), Howard-Jones <i>et al.</i> (2009),	Translation & back-translation	Not reported
Ruhaak&Cook, 2018	25-item neuromyth identification tool	Dekker <i>et al.</i> (2012), Howard-Jones <i>et al.</i> (2009), OECD (2002)	Not reported	Not reported
Ferrero <i>et al.</i> , 2016	32-item neuromyth identification tool	Dekker <i>et al.</i> (2012), Howard-Jones <i>et al.</i> (2009), OECD (2002)	Not reported	Not reported
Tardif <i>et al.</i> , 2015	3-item neuromyth data collection tool	Not reported	Not reported	Not reported
Dekker <i>et al.</i> , 2012	32-item neuromyth identification tool	Howard-Jones <i>et al.</i> (2009), OECD (2002)	Not reported	Not reported

## 2. METHOD

This research employed a mixed-methods approach, combining both quantitative and qualitative methodologies to develop a valid and reliable instrument for identifying teachers' neuromyth beliefs and their sources. The development of the questionnaire involved four stages: literature review, design, pilot testing, and language adaptation. During the literature review phase, a comprehensive review of existing studies and thematic analysis were conducted

to create an initial pool of questionnaire items and to validate the content of these items. Subsequently, in the design phase, the items were carefully crafted in terms of both content and presentation, setting the stage for the pilot phase. The pilot study served as a preliminary test to evaluate the characteristics of the questionnaire items. Finally, in the language adaptation phase, the questionnaire was translated to accommodate the linguistic diversity of the target audience, and its structural integrity was analyzed.

## 2.1. Analysis and Design Phase: Literature Analysis and Creation of Items

The analysis phase of the study was characterized by a comprehensive review and descriptive analysis. The research encompassed a thorough examination of most studies on neuromyths that were present within international databases. This comprehensive review aimed to identify prevalent neuromyths and to evaluate both formal and informal learning resources associated with them.

### 2.1.1. Literature review criteria

The literature review was meticulously structured based on criteria established by the research team. To ensure a comprehensive and systematic search, specific parameters were established for keywords, databases, areas of research, language, and year of publication. Keywords selected for the literature search included “neuromyths”, “neuromyth belief” and “neuromyths and learning”. The databases selected for review were those containing articles related to “education”, “educational sciences”, “educational research”, “psychology”, and “neuroscience”, specifically indexed in renowned academic platforms such as Web of Science (WOS), Elsevier, Springer, Sage, Taylor & Francis, Wiley, and Frontiers Media Sa. (SSCI, ESCI-SCI-E). The language of the research was limited to English. Given the nascent nature of the field of neuromyth and the resulting paucity of studies, we chose not to restrict the year of publication. A total of 141 studies were initially identified following the review process. During the planning and review process, three researchers conducted independent searches using the same keywords and obtained a similar number of studies. Five duplicate studies were removed from the review in the selection phase ( $k = 136$ ). Then, 39 studies that were not written in English or that were not research articles were excluded ( $k = 97$ ). Based on the subject-area criterion, an additional 35 studies unrelated to the fields of education, neuroscience, or psychology were removed ( $k = 62$ ). According to these criteria, the remaining studies were reclassified by content. This resulted in 49 articles associated with educational sciences and learning. We then examined the research aims, methodological designs, data collection instruments, and analyses of these selected studies in greater detail in terms of accuracy, clarity, and validity. Finally, the articles were reclassified according to their content, and 37 studies were analyzed (Adıgüzel et al., 2024). The detailed criteria are presented in Table 2.

**Table 2.** Literature review criteria.

Categories	Criteria
Keyword:	“Neuromyths”, “Neuromyths belief” “Neuromyths and learning”
Database:	Web of Science (WOS), Elsevier, Springer, Sage, Taylor & Francis, Wiley, Frontiers Media Sa
Field:	“education”, “educational sciences”, “educational research”, “psychology”, “neuroscience”
Research type:	Article
Language:	English
Year:	All years



### **2.1.2. Evaluation and creation of items**

To develop a comprehensive and up-to-date questionnaire, an extensive evaluation of data collection instruments from relevant studies was conducted. Each item from these studies underwent a rigorous evaluation by researchers and subject matter experts, focusing on similarity, accuracy, clarity, and validity. This process included correlating each item with its respective references and compiling them into a centralized item repository. A panel of six Ph.D.-level experts with international contributions in neuroscience, neuro education and education reviewed the items, excluding those that failed to meet established criteria. Theoretical frameworks from the literature were also considered to inform the development of neuromyth items. Discussions ensured the appropriateness of each statement, considering the target audience's characteristics. This meticulous process ensured the development of a robust and comprehensive questionnaire.

Following the analysis of the neuromyth items, an in-depth review of learning resources was conducted. The resources were categorized as either formal or informal to enable a more comprehensive analysis. Upon completion of these studies, neuromyth items and learning resources were identified and then structured into a questionnaire form.

## **2.2. Implementation Phase: Pilot Implementation**

In the second phase of the study, the draft instrument was piloted, and after the piloting, a general evaluation meeting was held with the participants, and the items in the item repository were evaluated.

### **2.2.1. Participants**

The pilot implementation was conducted with teachers employed at a private school. No specific sampling procedure was applied. Instead, participation was based on voluntary consent, and all teachers at the school were invited to participate in the study. Participants were required to be actively employed as teachers at the time of the study. Those who were not engaged in teaching were excluded. A total of 190 teachers participated in the pilot implementation of the developed instrument. Two teachers were excluded for providing blank answers, and five were excluded for failing to respond to most items. Consequently, the analyses were based on the responses of 183 teachers. Participants in the study were predominantly female, with 78% of teachers being women and 22% being men. Among the participating teachers, 42% had five years of experience or less, 27% had six to ten years, 13% had eleven to fifteen years, 7% had sixteen to twenty years, and 11% had twenty-one years or more.

### **2.2.2. Pilot implementation process**

During the pilot phase, the draft questionnaire designed to assess cognitive readiness and relevant learning resources was to be administered in a controlled setting. This draft questionnaire was administered to 190 teachers in a single session, with an average completion time of 20 minutes per respondent. Following the session, an analysis of the 39-item instrument was conducted. A follow-up meeting was convened to discuss the results of the analysis and to address any items that participants found overly technical or challenging.

### **2.2.3. Analysis of pilot implementation data**

In the present study, several statistical procedures were applied to ensure the validity, reliability, and cultural appropriateness of the measurement instrument. Subsequently, exploratory factor analysis (EFA) was employed to examine the construct validity of the instrument and to verify whether the empirical data supported the theoretically proposed two-factor structure ("Perceptions of the learning process" and "Perceptions of brain and intelligence characteristics"). The Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity were calculated to determine the suitability of the data for factor analysis. Factor loadings, eigenvalues, and scree plot inspection were used to confirm the dimensional structure. In addition, correlation analyses were performed to explore the

relationships between the identified factors and relevant study variables, providing further evidence for the instrument's construct validity. Internal consistency was evaluated using Cronbach's alpha coefficients. All analyses were conducted using SPSS, and a significance level of  $\alpha < .05$  was adopted.

### **2.3. Language Adaptation Phase: French, Turkish, Greek, Kazakh, Arabic, Malay, Chinese**

#### **2.3.1. Language adaptation protocol and briefing**

To ensure the global applicability of the developed questionnaire, a comprehensive selection of countries was identified for the implementation study. To facilitate this, language adaptation criteria were carefully established. Collaborations with scientists and translators from each target country were initiated to ensure accurate adaptation. The experts recruited for this adaptation process were guided by the following key phases:

**Preparation Phase:** This involved the formation of a dedicated translation team, followed by detailed briefings and specialized training to align the team with the study objectives.

**Translation and back-translation phase:** The questionnaire underwent a rigorous translation process, followed by back-translation to ensure fidelity to the original content. This phase also included careful comparison, editing, and review by a panel of experts before final implementation.

#### **2.3.2. Language adaptation process**

In this study, translating the data collection instrument into multiple languages was a critical step. Researchers from each target country were recruited on the basis of predefined criteria, and the English version of the instrument was distributed to them. These researchers played an integral role in the adaptation process, ensuring that the tool was modified according to the established guidelines. Through this collaborative effort, the data collection tool was prepared for global use in eight languages: French, English, Turkish, Greek, Kazakh, Arabic, Malay, and Chinese.

### **2.4. Validity and Reliability of The Study**

Within the scope of the research, ethical and implementation permissions were obtained in accordance with established ethical guidelines. These permissions were subsequently shared with all participating countries, ensuring a unified and compliant approach to the study.

#### **2.4.1. Validity**

To ensure the validity of the questionnaire, several steps were followed. First, a comprehensive literature review was conducted to inform the development of the questionnaire items. This review was essential for grounding the questionnaire in scientific evidence and incorporating the full scope of knowledge and findings from previous research.

The study's ethical validity was ensured by the voluntary participation of respondents. Additionally, the research received ethics committee approval, which was disseminated to administrators in each participating country, further affirming the study's adherence to ethical standards and guidelines.

The formulation of the questionnaire items and their translation into different languages were informed by expert consultation. Each item was supported by a literature review, thereby strengthening the content validity of the instrument. In addition, the inclusion of a wide range of learning resources, both formal and informal, ensured a comprehensive representation of participants' learning environments within the survey.

#### **2.4.2. Reliability**

A series of strategies was used to increase the reliability of the study. Initial pilot testing of the questionnaire was used to clarify any ambiguous items. To achieve consistent results across different country groups, maximum sampling variation and objective sample selection were

used. The language adaptation phase involved careful translation and back-translation of the questionnaire items to ensure linguistic accuracy and comprehensibility. The inclusion of both correct and incorrect items, along with the strategic use of the term “perception” rather than “neuromyth” was designed to minimize researcher-induced bias. To complement the neuromyth items, learning resources were developed through an extensive literature review and expert consultation, and the questionnaire was formally structured to reflect these inputs. The internal consistency of the questionnaires across languages was verified using the split-half method. In addition, inter-language correlation coefficients were calculated using data from a new application.

During the language adaptation process, responses collected through the online survey platform were analyzed. To prevent participants from searching for correct answers online, a 30-second time limit was imposed for each item. These measures collectively enhanced the questionnaire's reliability, improving both its internal consistency and temporal stability.

### 3. RESULTS

#### 3.1. Analysis and Design Phase: Literature Analysis and Development of Items

During the analysis and design phase of the study, a comprehensive, multifaceted questionnaire was developed to identify teachers' misconceptions about neuromyths and their associated learning resources in relation to the learning process and intelligence characteristics in eight languages. This development was preceded by a systematic review of the existing literature on neuromyths, which yielded an initial 50-item instrument that aligned with the research methodology's requirements.

Draft items were reviewed by a panel of six experts, each with at least a master's degree in neuroscience or education. Their collective insights resulted in a refined questionnaire that was narrowed down to 39 items selected for their relevance, informational accuracy, and alignment with the research objectives and phrasing style.

To authenticate the data collection tool and facilitate nuanced data analysis, the questionnaire was divided into two dimensions: “Perceptions of the learning process” and “Perceptions of brain and intelligence characteristics”. The format of the questionnaire was formalized to allow for detailed examination. Response categories such as “correct”, “incorrect” and “I can't answer because” were introduced for each neuromyth statement. Additional response options, “I don't have enough information to answer”, “There are important uncertainties about this issue”, “I've never heard of it” were included to clarify the reasons for unanswered items and to minimize data attrition.

A distinctive feature of the questionnaire is the inclusion of categories for both formal and informal sources that shape neuromyth beliefs. To thoroughly investigate the origins of these beliefs, 12 different sources were identified and incorporated into the data collection instrument, as shown in Figure 1.

Sources  Statements	Formal learning resources			Informal learning resources								NC
	S1. Undergraduate/graduate education	S2. Professional development programs	S7. Academic publications	S3. Professional experience	S4. Social media	S5. Websites	S6. Colleagues or friends	S8. Other publications	S9. The movies	S10. Television programs	S11. Advertisements	S12. Self-intuitions

**Figure 1.** Neuromyth learning resources questionnaire.

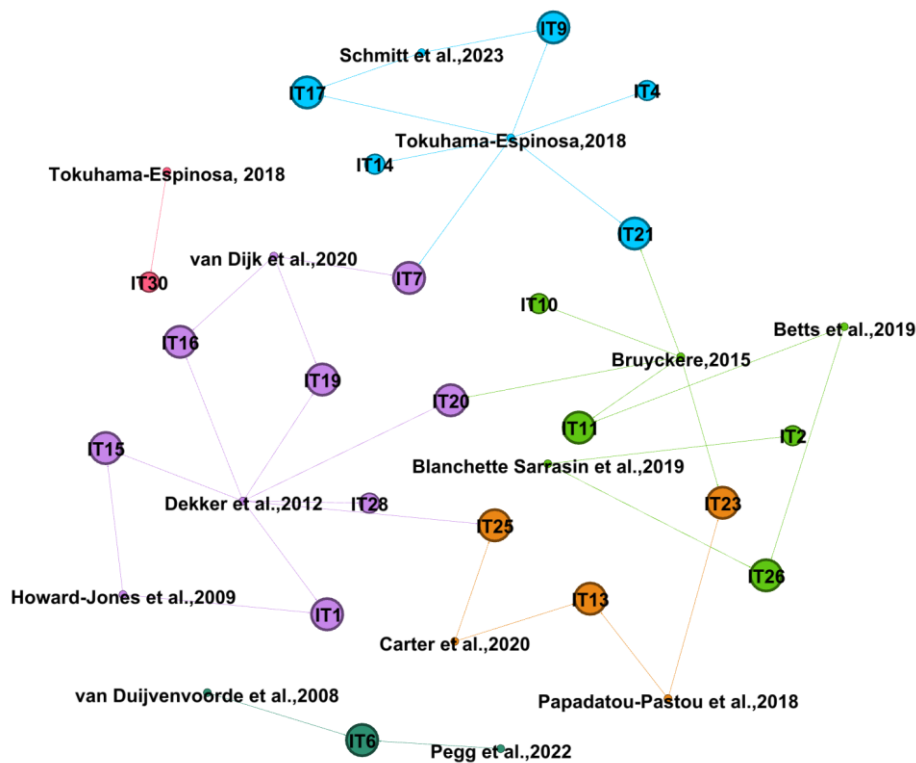
As seen in Figure 1, formal learning resources were grouped under three sources, and informal learning resources were grouped under eight sources because they were more diverse. Within this structure, “self-intuition” was structured outside of the formal and informal sources. This origin, which was included at the suggestion of experts in the field, was included in the categories as an independent learning resource formed by professional experience and affective learning characteristics.

To determine the neuromyths and learning resources, each neuromyth item was defined based on various references within a comprehensive literature review. Examples of these references are provided in Table 3.

**Table 3.** Scientific references of neuromyths items.

Items	References	Items	References
IT1	Dekker <i>et al.</i> , 2012	IT17	Schmitt <i>et al.</i> , 2023
IT4	Tokuhamma-Espinosa, 2018	IT19	van Dijk <i>et al.</i> , 2020
IT9	Schmitt <i>et al.</i> , 2023	IT20	Dekker <i>et al.</i> , 2012
IT10	De Bruyckere, 2015	IT21	De Bruyckere, 2015
IT11	De Bruyckere, 2015	IT23	De Bruyckere, 2015
IT13	Papadatou-Pastou <i>et al.</i> , 2018	IT25	Dekker <i>et al.</i> , 2012
IT14	Tokuhamma-Espinosa, 2018	IT26	Betts <i>et al.</i> , 2019
IT15	Howard-Jones <i>et al.</i> , 2009	IT28	Dekker <i>et al.</i> , 2012
IT16	Dekker <i>et al.</i> , 2012		

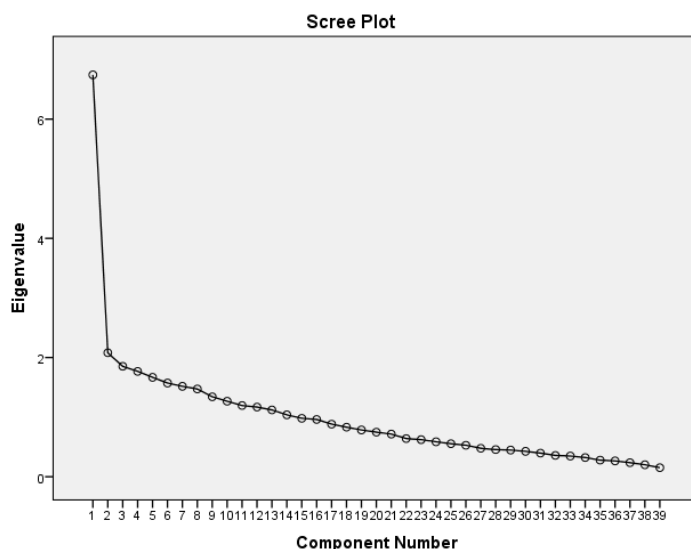
The data shown in Table 3 are also shown in Figure 2 with network analysis through the Gephi 0.10.1 program. This representation allows an understanding of the invocation of each item in the studies from which they were inspired.



**Figure 2.** Neuromyths network analysis.

### 3.2. Implementation Phase: Pilot Implementation

The questionnaire was structured around two dimensions that were theoretically predetermined: “Perceptions of the learning process” (ST1–ST19) and “Perceptions of brain and intelligence characteristics” (ST20–ST39). The items and the theoretical framework were designed from the outset according to this two-factor structure. A factor analysis was subsequently conducted to validate that the neuromyth items were appropriately categorized under these dimensions. Prior to conducting the Exploratory Factor Analysis (EFA) to examine the construct validity of the instrument, the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett’s test of sphericity were performed to assess the suitability of the data for factor analysis. The KMO value was 0.70, indicating an acceptable level of sampling adequacy. Bartlett’s test of sphericity yielded a statistically significant result ( $\chi^2(741) = 1754.95, p < .001$ ), confirming that the correlation matrix was factorable. The EFA was conducted using the Principal Component Analysis extraction method in SPSS. To enhance interpretability and achieve a simpler factor structure, the Equamax orthogonal rotation method was applied. Prior to conducting the factor analysis, the dataset was examined for both univariate and multivariate outliers. Initially, all item responses were screened for completeness and consistency. Furthermore, an examination of the inter-item correlation matrix was conducted on the dataset. All correlations between variables were below .90, indicating that multicollinearity was not a concern. The results of the analysis confirmed the hypothesized two-factor structure of the instrument (1) Perceptions of the Learning Process and (2) Perceptions of Brain and Intelligence Characteristics as theoretically proposed. This plot is illustrated in Figure 3.



**Figure 3.** Scree plot graph.

Following the initial exploratory factor analysis (EFA), the structure of the instrument was confirmed to align with the theoretically proposed two-factor model. To refine the scale, item factor loadings were examined in detail after applying the Equamax rotation method. Items exhibiting cross-loadings above .30 on both factors or loadings below the acceptable threshold of .40 were identified as problematic (Tabachnick & Fidell, 2013). Based on these statistical criteria, nine items (Items 2, 4, 8, 16, 25, 30, 31, 32, and 39) were systematically removed from the instrument in an iterative process. After each item removal, the factor analysis was rerun to ensure that the two-factor structure remained stable and that the overall model fit improved. This process resulted in a final 30-item instrument with clear and distinct loadings on the two dimensions (Perceptions of the Learning Process and Perceptions of Brain and Intelligence Characteristics), free from cross-loading items and demonstrating improved internal consistency. The rotated factor loadings of the retained items are presented in Table 4.



**Table 4.** Rotated component matrix.

Component			Component			Component		
Items	1	2	Items	1	2	Items	1	2
IT35	.646	.162	IT28	.449		IT20	.213	.587
IT22	.624		IT17	.406	.109	IT13		.556
IT38	.587	.107	IT14	.370		IT12	.251	.496
IT36	.555	.180	IT19	.353	.205	IT18		.396
IT33	.553		IT23	.331		IT10	.154	.378
IT7	.549		IT29	.328	.316	IT27	.105	.301
IT24	.531	.188	IT21	.324	.265	IT15	.237	.290
IT37	.527	.165	IT26	.307	.248	IT5		.230
IT6	.459		IT11	.185	.645	IT3		.220
IT34	.451	.189	IT9	.260	.632	IT1	-.111	.147

The analyses revealed that the empirical factor structure largely aligned with the anticipated two dimensions. Specifically, items IT6, IT7, IT14, IT17, and IT19 loaded on the "Perceptions of Brain and Intelligence Characteristics" factor, and items ST20 and ST27 loaded on the "Perceptions of the Learning Process" factor. However, several items (IT1, IT3, IT5, IT10, IT15, IT18, and IT27) had factor loadings below the recommended threshold of .40, indicating an insufficient association with their respective constructs (Hair *et al.*, 2019). Iterative examinations of the rotated component matrix also identified nine items (2, 4, 8, 16, 25, 30, 31, 32, and 39) with low loadings or substantial cross-loadings ( $\geq .30$  on both factors). These items were removed to improve factorial clarity. After systematically eliminating these items and rerunning the EFA after each removal, a final 30-item instrument emerged, demonstrating a clear two-factor solution free from cross-loading items. The initial 39-item scale had a Cronbach's alpha coefficient of .88, indicating high internal consistency. This remained strong ( $\alpha = .85$ ) for the refined 30-item version. The final questionnaire included 21 false and nine true items, and the reduced set of items showed improved interpretability and construct validity.

After all revisions were made, the final data collection instrument included 30 items. Twenty-one of these items were formulated to be incorrect (IT1, IT2, IT4, IT6, IT7, IT9, IT10, IT11, IT13, IT14, IT15, IT16, IT17, IT19, IT20, IT21, IT23, IT25, IT26, IT28, IT30), and nine were formulated to be correct (IT3, IT5, IT8, IT12, IT18, IT22, IT24, IT27, IT29). To minimize response bias and maintain measurement validity, the correct and incorrect items were intermixed throughout the questionnaire rather than being grouped by type, as illustrated in [Table 5](#). In addition, the complete version of the neuromyth diagnostic instrument is presented in the Appendices in eight languages: [Appendix 1](#) (English), [Appendix 2](#) (Turkish), [Appendix 3](#) (Arabic), [Appendix 4](#) (Chinese), [Appendix 5](#) (French), [Appendix 6](#) (Greek), [Appendix 7](#) (Kazakh), and [Appendix 8](#) (Malay).

**Table 5.** Neuromyth items.

No	Items	Correct/ Incorrect
M1	Individuals learn better when they receive information in alignment with in their dominant learning styles (examples: visual, auditory, kinesthetic etc.)	Incorrect
M2	The dominant intelligence profile of learners (examples: mathematical, verbal, spatial) must be considered in teaching	Incorrect
M3	In the learning process, the mind associates new information with previous knowledge	Correct
M4	Different parts of the brain operate independently during the learning process	Incorrect

**Table 5.** *Continues.*

M5	Learning occurs through changes in synaptic connections between neurons in the brain	Correct
M6	Learning is a purely cognitive skill, not emotional	Incorrect
M7	Learning takes place independent from individuals' learning backgrounds	Incorrect
M8	Some mental processes (experience, learning) repeated over a long period of time can change the structure and function of some areas of the brain	Correct
M9	Individuals can learn new information even while in a state of sleep	Incorrect
M10	Humans are good multitaskers	Incorrect
M11	The fact that some people are more “right-brained” and others are more “left-brained”, helps explain the differences in how we learn	Incorrect
M12	Individuals learn better when course content is presented in short sessions or modules	Correct
M13	There are specific periods in childhood after which certain things can no longer be learned	Incorrect
M14	Memorization has no impact on the learning process	Incorrect
M15	Environments that provide a larger amount of stimuli improve the brains of pre-school children	Incorrect
M16	Mental capacity is hereditary and cannot be changed by the environment or experience	Incorrect
M17	Listening to classical music improves mental capacity	Incorrect
M18	When a part of the brain is damaged, other parts can take over its function	Correct
M19	Short periods of coordination exercises can improve brain function (for example, touching your right ankle with your left hand and vice versa)	Incorrect
M20	We use only 10% of our brain	Incorrect
M21	Individuals with larger brains are smarter	Incorrect
M22	The brain continues to generate new connections throughout an individual's life	Correct
M23	Male and female brains are designed for different types of skills	Incorrect
M24	The brain remains active 24 hours a day	Correct
M25	Supplements such as Omega-3 and Omega-6 have a positive effect on academic achievement	Incorrect
M26	Brain development is complete by the time children reach the end of puberty	Incorrect
M27	The normal development of the human brain involves the birth and death of brain cells	Correct
M28	The brain shuts down during sleep	Incorrect
M29	On average, males have bigger brains than females	Correct
M30	Humans are born with all the neurons they will have in their lifetime	Incorrect

### 3.3. Language Adaptation: French, Turkish, Greek, Kazakh, Arabic, Malay, Chinese

To facilitate the translation of the data collection tool into multiple languages, the researchers shared the English version with partner countries and requested translations into their respective languages. During the translation process, an adaptation protocol was also prepared and sent with the measurement tool. The following steps were followed in the process of language adaptation of the measurement tool.

#### 3.3.1. Organization phase

Establishing the Translation Team: The first step in the multinational research project was to develop a measurement tool. Teams were assembled from each participating country that were proficient in the target languages. An English version of the Neuromyth Questionnaire was

distributed via e-mail, which led to the creation of specialized translation teams composed of experts in both the subject matter and the respective languages. These teams were then responsible for adapting the data collection tool to their cultural and linguistic contexts.

**Information and Training:** The translation teams were informed about the purpose and content of the questionnaire and about neuromyths. The translation process highlighted the crucial importance of linguistic nuance and cultural relevance. In addition, teams were instructed to pay close attention to terminology used in educational and neuroscientific disciplines in order to maintain the integrity of the questionnaire across languages.

### 3.3.2. Translation and back translation process

**Translation:** Two independent linguists translated the original questionnaire into the respective target languages.

**Back translation:** A different pair of independent linguists then back-translated these versions into the original language.

**Comparison and editing:** The original questionnaire and the back-translated versions were meticulously compared. Discrepancies were addressed, leading to refinements in both the translated and back-translated texts.

### 3.3.3. Expert panel and pretest

**Expert panel:** Translated versions of the questionnaire were e-mailed to the researchers. The translations were subject to rigorous review process, involving both expert language specialists and advanced computational tools, to detect and correct any errors or deficiencies. As a result, the research data collection instrument was accurately refined and made available in Turkish, Greek, Kazakh, Arabic, Malay, Chinese, English, and French.

**Pilot Application:** Once the data collection instrument was finalized in eight languages, it was integrated into the online survey platform “Interceptum”. The survey links were subsequently distributed to the participating countries, where teachers were recruited to facilitate the collection of empirical data. Following data collection, a subset of 50 participants per language was extracted to form a new, representative dataset. This data set was used to calculate the internal consistency coefficient of the instrument by assessing the correlation between its two halves. The calculated consistency coefficients for each language are shown in [Table 6](#).

**Table 6.** Internal consistency coefficients.

Correlation	Language	French ( <i>f</i> = 50)	Turkish ( <i>f</i> = 50)	English ( <i>f</i> = 50)	Greek ( <i>f</i> = 50)	Kazakh ( <i>f</i> = 50)	Malay ( <i>f</i> = 50)	Arabic ( <i>f</i> = 50)	Chinese ( <i>f</i> = 50)
Spearman-Brown Coefficient	Equal Length	.74	.89	.93	.81	.91	.82	.85	.78
	Unequal Length	.74	.89	.93	.81	.91	.82	.85	.78
Guttman Split Half Coefficient		.74	.78	.92	.75	.89	.74	.75	.68

### 3.3.4. Validity and reliability analyses

To evaluate the psychometric quality of the instrument across all language versions, both reliability and validity analyses were conducted on the pilot data. The reliability assessment focused on internal consistency, while the validity evaluation emphasized cross-linguistic construct equivalence.

Internal consistency was examined using the Spearman–Brown split-half coefficient, Guttman split-half coefficient, and Cronbach's alpha for reliability. For all eight language versions (French, Turkish, English, Greek, Kazakh, Malay, Arabic, and Chinese), the coefficients exceeded the commonly accepted threshold of .70 (Nunnally & Bernstein, 1994). Specifically,

the Spearman–Brown coefficients ranged from .74 to .93 and the Guttman split-half coefficients ranged from .68 to .92 (see Table 6). These results indicate that the items within each language version consistently measure the intended constructs, demonstrating satisfactory to excellent internal consistency. The English version exhibited the highest internal consistency ( $\alpha = .93$ ), while all other languages remained well above the acceptable range.

Construct validity across languages was assessed through interlanguage correlation analyses using Spearman's rho. High and statistically significant correlations were found between all language pairs ( $r = .98-.99$ ,  $p < .001$ ; see Table 7). These extremely high coefficients provide strong evidence that the instrument consistently measures the same underlying construct across diverse linguistic and cultural contexts. The findings suggest that the translation, back-translation, and expert panel review processes ensured semantic and conceptual equivalence among the different language versions. This cross-linguistic consistency confirms that the adapted versions maintain the factorial structure and interpretive meaning of the original instrument. In summary, the multilingual adaptations of the neuromyth identification tool demonstrated high internal consistency and provided substantial evidence of cross-linguistic construct validity. These results confirm the tool's methodological robustness and support its use in rigorous, comparative studies of neuromyth beliefs in different linguistic and cultural contexts.

**Table 7.** Correlation coefficients.

Language		French ( $f = 50$ )	Turkish ( $f = 50$ )	English ( $f = 50$ )	Greek ( $f = 50$ )	Kazakh ( $f = 50$ )	Malay ( $f = 50$ )	Arabic ( $f = 50$ )	Chinese ( $f = 50$ )
French	Correlation Coefficient	1.000	.991**	.989**	.990**	.980**	.987**	.989**	.983**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000
Turkish	Correlation Coefficient	.991**	1.000	.990**	.988**	.981**	.988**	.990**	.984**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000
English	Correlation Coefficient	.989**	.990**	1.000	.991**	.986**	.989**	.992**	.993**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000
Greek	Correlation Coefficient	.990**	.988**	.991**	1.000	.987**	.990**	.989**	.989**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000	.000
Kazakh	Correlation Coefficient	.980**	.981**	.986**	.987**	1.000	.987**	.985**	.989**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000
Malay	Correlation Coefficient	.987**	.988**	.989**	.990**	.987**	1.000	.990**	.991**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000
Arabic	Correlation Coefficient	.989**	.990**	.992**	.989**	.985**	.990**	1.000	.988**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000		.000
Chinese	Correlation Coefficient	.983**	.984**	.993**	.989**	.989**	.991**	.988**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	

\*\* Correlation is significant at the .01 level (2-tailed).

## 4. CONCLUSION

This study developed and validated a multilingual neuromyth identification tool to measure the prevalence and origins of neuromyth beliefs among teachers. Psychometric analyses revealed that the final 30-item scale exhibited strong internal consistency, a consistent two-factor structure, and substantial cross-linguistic construct validity across eight language versions. These findings have important methodological, theoretical, and practical implications.

The exploratory factor analysis (EFA) confirmed the theoretically anticipated two-factor structure, “Perceptions of the Learning Process” and “Perceptions of Brain and Intelligence Characteristics” with items exhibiting clear and distinct loadings. Through an iterative process, problematic items with factor loadings below .40 or cross-loadings of at least .30 were systematically removed, resulting in a 30-item instrument. This rigorous approach to eliminating items ensured factorial clarity and internal consistency (Cronbach’s  $\alpha = .85$ ). The high factor loadings of the retained items (most exceeding .50) indicate strong alignment with the intended constructs. Items that were removed from the scale often represented items that were subject to differences in cultural interpretation or conceptual ambiguity during the translation process.

The present instrument offers several methodological innovations compared to widely used neuromyth questionnaires, such as the Dekker *et al.* (2012) scale and the Neuroscience Against Neuromyths Questionnaire (Tovazzi *et al.*, 2020). First, it was designed from the beginning to be multilingual and culturally adaptable. It encompasses eight languages that represent diverse geographic and educational contexts. Second, the scale incorporates a balanced mix of true and false items to reduce acquiescence bias and provide a more accurate assessment of knowledge. Third, the tool uniquely incorporates an assessment of the sources from which participants acquire knowledge about brain function, addressing a notable gap in literature.

This study addresses a critical gap in the neuromyth literature by developing and validating a multilingual instrument with strong psychometric properties. Previous research has largely relied on monolingual instruments with limited cross-cultural testing and has often neglected rigorous adaptation procedures. This new tool allows for reliable cross-cultural comparisons of neuromyth beliefs, thus broadening the scope of educational neuroscience research.

The findings of this study have significant implications for teacher education, curriculum development, and educational policy. The high prevalence of misconceptions regarding certain items emphasizes the urgent necessity of implementing focused professional development programs to enhance teachers’ neuroscience literacy. The validated instrument developed in this study can serve as a reliable diagnostic tool in pre-service and in-service teacher education. It enables educators and policymakers to identify and address the most persistent neuromyths through evidence-based interventions. The two-factor structure, which captures misconceptions about learning processes and brain/intelligence characteristics, suggests that teacher training curricula should integrate neuroscience content that directly addresses these misconceptions.

Beyond teacher education, the instrument’s multilingual design and robust psychometric properties allow for valid cross-national comparisons of neuromyth prevalence in diverse cultural and linguistic contexts. Such comparative data can guide policymakers in developing context-specific strategies while contributing to a global framework for addressing misconceptions in education. Additionally, including items that assess knowledge sources provides a unique opportunity to investigate how formal and informal information channels influence the persistence of neuromyths. This offers both theoretical insight and practical guidance for designing more effective educational interventions.

### 4.1. Limitations

Despite its methodological thoroughness and multilingual scope, this study has several restrictions that should be recognized. First, although rigorous translation and back-translation procedures were employed to ensure linguistic accuracy, the complexity of the neuroscience



concepts may have introduced subtle differences in meaning across languages that could affect respondents' interpretations. Second, the language adaptation process was conducted with relatively small groups of participants in each target language. Expanding this stage to include larger, more diverse samples would strengthen the validity of cross-linguistic comparisons. Second, adapting the instrument into eight languages represents a significant improvement over existing tool. However, adapting it to additional languages and cultural contexts would increase its global applicability and ensure that it captures a broader range of educational settings. Third, while the pilot and validation samples were diverse, they were limited in size. In some cases, they were drawn from convenience samples within specific educational institutions. This may restrict the generalizability of the findings, so future studies should test the instrument on larger, more representative populations.

### Acknowledgments

The authors wish to extend their gratitude to all research partners from participating countries for their invaluable collaboration in the development and adaptation of the instrument. The authors also acknowledge using AI-assisted language tools to proofread and refine the language of this manuscript. These tools were used to improve the clarity and coherence of the text without altering its intellectual content or analytical interpretations.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Ethics Committee of Anadolu University, 04.12.2023-649986.

### Contribution of Authors

**Oktaý Cem Adıgüzel:** Conceptualization, Investigation, Resources, Visualization, Software, Formal analysis, and Writing-original draft. **Sibel Küçük kayhan:** Conceptualization, Methodology, Supervision, and Validation. **Patrice Potvin:** Supervision, Validation, Critical review and editing. **Derya Atik-Kara:** Conceptualization, Formal analysis, Writing-original draft, and Visualization.

### Orcid

Oktaý Cem Adıgüzel  <https://orcid.org/0000-0002-7985-4871>

Sibel Küçük kayhan  <https://orcid.org/0000-0002-3035-9220>

Patrice Potvin  <https://orcid.org/0000-0002-1623-2362>

Derya Atik-Kara  <https://orcid.org/0000-0002-6890-030X>

### REFERENCES

- Adıgüzel, C., Atik-Kara, D., & Küçük kayhan, S. (2024). Misconceptions of the learning and teaching process: Neuromyths and their formal and informal sources introduction. *Türk Psikoloji Yazıları*, 27(53), 68-70. <https://doi.org/10.31828/tpy13019966>
- Betts, K., Miller, M., Tokuhama-Espinosa, T., Shewokis, P., Anderson, A., Borja, C., Galoyan, T., Delaney, B., Eigenauer, J., & Dekker, S. (2019). *International report: Neuromyths and evidence-based practices in higher education*. Online Learning Consortium.
- Bissessar, S., & Youssef, F.F. (2021). A cross-sectional study of neuromyths among teachers in a Caribbean nation. *Trends in Neuroscience and Education*, 23, Article 100155. <https://doi.org/10.1016/j.tine.2021.100155>
- Carter, M., van Bergen, P., Stephenson, J., Newall, C., & Sweller, N. (2020). Prevalence, predictors, and sources of information regarding neuromyths in an Australian cohort of preservice teachers. *Australian Journal of Teacher Education*, 45(10), 95-113. <https://doi.org/10.14221/ajte.2020v45n10.6>

- Ching, F.N.Y., So, W.W.M., Lo, S.K., & Wong, S.W.H. (2020). Preservice teachers' neuroscience literacy and perceptions of neuroscience in education: Implications for teacher education. *Trends in Neuroscience and Education*, 21, Article 100144. <https://doi.org/10.1016/j.tine.2020.100144>
- Craig, H.L., Wilcox, G., Makarenko, E.M., & MacMaster, F.P. (2021). Continued educational neuromyth belief in pre- and in-service teachers: A call for de-implementation action for school psychologists. *Canadian Journal of School Psychology*, 36(2), 127-141. <https://doi.org/10.1177/0829573520979605>
- Deans, C., & Larsen, E. (2022). Brain-based Learning: Beliefs and Practice in one Australian Primary School Implementing a Neuroscience Pedagogical Framework. *Australian Journal of Teacher Education*, 47(10), 18-38. <https://doi.org/10.14221/ajte.2022v47n10.2>
- De Bruyckere, P., Kirschner, P.A., & Hulshof, C.D. (2015). *Urban myths about learning and education*. Academic Press.
- Deibl, I., & Zumbach, J. (2023). Pre-service teachers' beliefs about neuroscience and education-Do freshmen and advanced students differ in their ability to identify myths?. *Psychology Learning & Teaching*, 22(1), 74-93. <https://doi.org/10.1177/14757257221146649>
- Dekker, S., Lee, N.C., Howard-Jones, P., & Jolles, J. (2012). Neuromyths in education: Prevalence and predictors of misconceptions among teachers. *Frontiers in Psychology*, 3, Article 429. <https://doi.org/10.3389/fpsyg.2012.00429>
- Ferrero, M., Garaizar, P., & Vadillo, M.A. (2016). Neuromyths in education: Prevalence among Spanish teachers and an exploration of cross-cultural variation. *Frontiers in Human Neuroscience*, 10, Article 496. <http://dx.doi.org/10.3389/fnhum.2016.00496>
- Hair, J.F., Babin, B.J., Anderson, R.E., & Black, W.C. (2019). *Multivariate Data Analysis* (8<sup>th</sup> ed.). Pearson Prentice.
- Howard-Jones, P.A. (2014). Neuroscience and education: Myths and messages. *Nature Reviews Neuroscience*, 15(12), 817-824. <https://doi.org/10.1038/nrn3817>
- Jeyavel, S., Pandey, V., Rajkumar, E., & Lakshmana, G. (2022). Neuromyths in education: Prevalence among south Indian school teachers. *Frontiers in Education*, 7, Article 781735. <https://doi.org/10.3389/educ.2022.781735>
- Krammer, G., Vogel, S.E., Yardimci, T., & Grabner, R.H. (2019). Neuromythen sind zu Beginn des Lehramtsstudiums prävalent und unabhängig vom wissen über das menschliche gehirn [Neuromyths are prevalent at the beginning of teacher education programs and independent of knowledge about the human brain]. *Zeitschrift für Bildungsforschung*, 9, 221-246. <https://doi.org/10.1007/s35834-019-00238-2>
- Krammer, G., Vogel, S.E., & Grabner, R.H. (2020). Believing in neuromyths makes neither a bad nor good student-teacher: The relationship between neuromyths and academic achievement in teacher education. *Mind, Brain, and Education*, 15(1), 54-60. <https://doi.org/10.1111/mbe.12266>
- Macdonald, K., Germine, L., Anderson, A., Christodoulou, J., & McGrath, L.M. (2017). Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in Psychology*, 8, Article 1314. <https://doi.org/10.3389/fpsyg.2017.01314>
- McMahon, K., Yeh, C.S.H., & Etchells, P.J. (2019). The impact of a modified initial teacher education on challenging trainees' understanding of neuromyths. *Mind, Brain, and Education*, 13(4), 288-297. <https://doi.org/10.1111/mbe.12219>
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3<sup>rd</sup> ed.). McGraw-Hill.
- Organisation for Economic Cooperation and Development. (2002). *Understanding the brain: Towards a new learning science*. OECD.

- Papadatou-Pastou, M., Gritzali, M., & Barrable, A. (2018). The learning styles Educational neuromyth: lack of agreement between teachers' judgments, Self-Assessment, and students' intelligence. *Frontiers in Education*, 3, Article 105. <https://doi.org/10.3389/feduc.2018.00105>
- Ruhaak, A.E., & Cook, B.G. (2018). The prevalence of educational neuromyths among pre-service special education teachers. *Mind, Brain, and Education*, 12(3), 155-161. <https://doi.org/10.1111/mbe.12181>
- Ruiz-Martin, H., Portero-Tresserra, M., Martínez-Molina, A., & Ferrero, M. (2022). Tenacious educational neuromyths: Prevalence among teachers and an intervention. *Trends in Neuroscience and Education*, 29, Article 100192. <https://doi.org/10.1016/j.tine.2022.100192>
- Sazaka, L.S.R., Hermida, M.J., & Ekuni, R. (2024). Where did pre-service teachers, teachers, and the general public learn neuromyths? Insights to support teacher training. *Trends in Neuroscience and Education*, 36, Article 100235. <https://doi.org/10.1016/j.tine.2024.100235>
- Schmitt, A., Wollschläger, R., Blanchette Sarrasin, J., Masson, S., Fischbach, A., & Schiltz, C. (2023). Neuromyths and knowledge about intellectual giftedness in a highly educated multilingual country. *Frontiers in Psychology*, 14, Article 1252239. <https://doi.org/10.3389/fpsyg.2023.1252239>
- Simoes, E., Foz, A., Petinati, F., Marques, A., Sato, J., Lepski, G., & Arévalo, A. (2022). Neuroscience knowledge and endorsement of neuromyths among educators: What is the scenario in Brazil? *Brain Sciences*, 12(6), Article 734. <https://doi.org/10.3390/brain-sci12060734>
- Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6<sup>th</sup> ed.). Pearson.
- Tardif, E., Doudin, P.A., & Meylan, N. (2015). Neuromyths among teachers and student teachers. *Mind, brain, and Education*, 9(1), 50-59. <https://doi.org/10.1111/mbe.12070>
- Tokuhamma-Espinosa, T. (2018). *Neuromyths: Debunking false ideas about the brain*. WW Norton & Company
- Tovazzi, A., Giovannini, S., & Basso, D. (2020). A new method for evaluating knowledge, beliefs, and neuromyths about the mind and brain among Italian teachers. *Mind, Brain, and Education*, 14(2), 187-198. <https://doi.org/10.1111/mbe.12249>
- van Elk, M. (2019). Socio-cognitive biases are associated to belief in neuromyths and cognitive enhancement: A pre-registered study. *Personality and Individual Differences*, 147, 28-32. <https://doi.org/10.1016/j.paid.2019.04.014>
- van Dijk, W., & Lane, H.B. (2020). The brain and the US education system: Perpetuation of neuromyths. *Exceptionality*, 28(1), 16-29. <https://doi.org/10.1080/09362835.2018.1480954>
- Vig, J., Révész, L., Kaj, M., Kälbli, K., Svraka, B., Révész-Kiszela, K., & Csányi, T. (2023). The prevalence of educational neuromyths among Hungarian pre-service teachers. *Journal of Intelligence*, 11(2), 31. <https://doi.org/10.3390/jintelligence11020031>
- Zhang, R., Jiang, Y., Dang, B., & Zhou, A. (2019, February). Neuromyths in Chinese classrooms: Evidence from headmasters in an underdeveloped region of China. *Frontiers in Education*, 4, Article 8. <https://doi.org/10.3389/feduc.2019.00008>

## APPENDICES

*Note on Item Classification.* The final version of the neuromyth diagnostic instrument comprises 30 items. Of these, 21 are incorrect items representing neuromyths (ST1, ST2, ST4, ST6, ST7, ST9, ST10, ST11, ST13, ST14, ST15, ST16, ST17, ST19, ST20, ST21, ST23, ST25, ST26, ST28, ST30), while 9 are correct items (ST3, ST5, ST8, ST12, ST18, ST22, ST24, ST27, ST29).

### Appendix 1. Neuromyths items learning process and intelligence characteristics: English version.

No	Items	Correct	Incorrect	For each item, please indicate which sources influenced your answers (You can check more than one answer)											
				1. Undergraduate/graduate education	2. Professional development programs	3. My professional experience	4. Social media (Twitter (X), Instagram, etc.)	5. Websites	6. My colleagues or friends	7. Academic publications	8. Other publications such as books, etc.	9. The movies	10. Television programs	11. Advertisements	12. My intuitions
M1	Individuals learn better when they receive information in alignment with in their dominant learning styles (examples: visual, auditory, kinesthetic etc.)			1	2	3	4	5	6	7	8	9	10	11	12
M2	The dominant intelligence profile of learners (examples: mathematical, verbal, spatial) must be considered in teaching			1	2	3	4	5	6	7	8	9	10	11	12
M3	In the learning process, the mind associates new information with previous knowledge			1	2	3	4	5	6	7	8	9	10	11	12
M4	Different parts of the brain operate independently during the learning process			1	2	3	4	5	6	7	8	9	10	11	12
M5	Learning occurs through changes in synaptic connections between neurons in the brain			1	2	3	4	5	6	7	8	9	10	11	12
M6	Learning is a purely cognitive skill, not emotional			1	2	3	4	5	6	7	8	9	10	11	12
M7	Learning takes place independent from individuals' learning backgrounds			1	2	3	4	5	6	7	8	9	10	11	12
M8	Some mental processes (experience, learning) repeated over a long period of time can change the structure and function of some areas of the brain			1	2	3	4	5	6	7	8	9	10	11	12
M9	Individuals can learn new information even while in a state of sleep			1	2	3	4	5	6	7	8	9	10	11	12
M10	Humans are good multitaskers			1	2	3	4	5	6	7	8	9	10	11	12
M11	The fact that some people are more "right-brained" and others are more "left-brained", helps explain the differences in how we learn			1	2	3	4	5	6	7	8	9	10	11	12
M12	Individuals learn better when course content is presented in short sessions or modules			1	2	3	4	5	6	7	8	9	10	11	12
M13	There are specific periods in childhood after which certain things can no longer be learned			1	2	3	4	5	6	7	8	9	10	11	12
M14	Memorization has no impact on the learning process			1	2	3	4	5	6	7	8	9	10	11	12
M15	Environments that provide a larger amount of stimuli improve the brains of pre-school children			1	2	3	4	5	6	7	8	9	10	11	12
M16	Mental capacity is hereditary and cannot be changed by the environment or experience			1	2	3	4	5	6	7	8	9	10	11	12
M17	Listening to classical music improves mental capacity			1	2	3	4	5	6	7	8	9	10	11	12
M18	When a part of the brain is damaged, other parts can take over its function			1	2	3	4	5	6	7	8	9	10	11	12
M19	Short periods of coordination exercises can improve brain function (for example, touching your right ankle with your left hand and vice versa)			1	2	3	4	5	6	7	8	9	10	11	12
M20	We use only 10% of our brain			1	2	3	4	5	6	7	8	9	10	11	12
M21	Individuals with larger brains are smarter			1	2	3	4	5	6	7	8	9	10	11	12
M22	The brain continues to generate new connections throughout an individual's life			1	2	3	4	5	6	7	8	9	10	11	12
M23	Male and female brains are designed for different types of skills			1	2	3	4	5	6	7	8	9	10	11	12
M24	The brain remains active 24 hours a day			1	2	3	4	5	6	7	8	9	10	11	12
M25	Supplements such as Omega-3 and Omega-6 have a positive effect on academic achievement			1	2	3	4	5	6	7	8	9	10	11	12
M26	Brain development is complete by the time children reach the end of puberty			1	2	3	4	5	6	7	8	9	10	11	12
M27	The normal development of the human brain involves the birth and death of brain cells			1	2	3	4	5	6	7	8	9	10	11	12
M28	The brain shuts down during sleep			1	2	3	4	5	6	7	8	9	10	11	12
M29	On average, males have bigger brains than females			1	2	3	4	5	6	7	8	9	10	11	12
M30	Humans are born with all the neurons they will have in their lifetime			1	2	3	4	5	6	7	8	9	10	11	12

**Appendix 2.** Neuromyths items learning process and intelligence characteristics: Turkish version.

No	Maddeler	Doğru Yanıt	Lütfen her bir madde için, yanıtınızı verirken hangi kaynaklardan etkilendiğinizi belirtiniz (Birden fazla yanıtı işaretleyebilirsiniz)											
			1. Lisans/lisansüstü eğitimlerden	2. Mesleki gelişim programlarından	3. Mesleki deneyimlerinden	4. Sosyal medya uygulamalarından (Twitter	5. Diğer internet sitelerinden	6. Meslektaşlarından veya arkadaşlarından	7. Akademik yayınlarından	8. Kitaplar, dergiler, popüler yayınlar gibi	9. Filmlerden	10. Televizyon programlarından	11. Reklamlardan	12. Mantıklı bulduğum için
M1	Öğrenciler, öğrenme stillerine uygun ortamlarda daha etkili öğrenirler (görsel, işitsel, kinestetik vb.)		1	2	3	4	5	6	7	8	9	10	11	12
M2	Öğretim sürecinde, öğrencilerin baskın olan zekâ türleri (matematiksel, sözel, uzamsal vb.) dikkate alınmalıdır		1	2	3	4	5	6	7	8	9	10	11	12
M3	Öğrenme sürecinde zihin yeni bilgilerle önceki bilgileri ilişkilendirir		1	2	3	4	5	6	7	8	9	10	11	12
M4	Öğrenme sürecinde beynin bölümleri birbirinden ayrı çalışır		1	2	3	4	5	6	7	8	9	10	11	12
M5	Öğrenme, beyindeki nöronlar arasındaki sinaptik bağlantılarda meydana gelen değişiklikler yoluyla gerçekleşir		1	2	3	4	5	6	7	8	9	10	11	12
M6	Öğrenme süreci duygusal değil, tamamıyla bilişsel bir özelliktir		1	2	3	4	5	6	7	8	9	10	11	12
M7	Öğrenme, bireylerin öğrenme geçmişlerinden bağımsız olarak gerçekleşir		1	2	3	4	5	6	7	8	9	10	11	12
M8	Uzun süreli tekrarlanan bazı zihinsel süreçler (deneyim, öğrenme vb.), beynin bazı bölgelerinin yapısını ve işlevini değiştirebilir		1	2	3	4	5	6	7	8	9	10	11	12
M9	Bireyler uykudayken yeni bilgiler öğrenebilir		1	2	3	4	5	6	7	8	9	10	11	12
M10	İnsanlar çoklu görevlerde başarılıdır		1	2	3	4	5	6	7	8	9	10	11	12
M11	Bazı insanların "sağ beyinli", bazılarının ise "sol beyinli" olması öğrenme özelliklerimizdeki farklılıkları açıklamaya yardımcı olur		1	2	3	4	5	6	7	8	9	10	11	12
M12	Bireyler, ders içeriği kısa oturumlar veya modüller halinde sunulduğunda daha iyi öğrenirler		1	2	3	4	5	6	7	8	9	10	11	12
M13	Çocuklukta belirli dönemlerden sonra bazı şeyler artık öğrenilemez		1	2	3	4	5	6	7	8	9	10	11	12
M14	Ezberlemenin öğrenme süreci üzerinde etkisi yoktur		1	2	3	4	5	6	7	8	9	10	11	12
M15	Zengin uyاریcı ortamlar okul öncesi dönemde çocukların beyinlerini daha iyi geliştirir		1	2	3	4	5	6	7	8	9	10	11	12
M16	Zihinsel kapasite kalıtsaldır ve çevresel faktörler veya öğrenme deneyimleri ile sonradan değiştirilemez		1	2	3	4	5	6	7	8	9	10	11	12
M17	Klasik müzik dinlemek zihinsel kapasiteyi geliştirir		1	2	3	4	5	6	7	8	9	10	11	12
M18	Beyin bölgesinin bir bölümü hasar gördüğünde beynin diğer bölgeleri bu işlevi üstlenebilir		1	2	3	4	5	6	7	8	9	10	11	12
M19	Kısa süreli koordinasyon egzersizleri beyin fonksiyonlarını geliştirebilir (örneğin, sol el ile sağ ayak bileğine dokunmak veya tam tersi)		1	2	3	4	5	6	7	8	9	10	11	12
M20	Beynimizin sadece %10'unu kullanırız		1	2	3	4	5	6	7	8	9	10	11	12
M21	Beyni büyük olan insanlar daha zeki olurlar		1	2	3	4	5	6	7	8	9	10	11	12
M22	Beyinde yeni bağlantıların üretimi yaşam boyu devam eder		1	2	3	4	5	6	7	8	9	10	11	12
M23	Erkek ve kadın beyinleri farklı türde beceriler için tasarlanmıştır		1	2	3	4	5	6	7	8	9	10	11	12
M24	Beynimiz günde 24 saat çalışmayı sürdürür		1	2	3	4	5	6	7	8	9	10	11	12
M25	Omega-3 ve Omega-6 gibi takviyelerin akademik başarı üzerinde olumlu etkisi vardır		1	2	3	4	5	6	7	8	9	10	11	12
M26	Bir kişi ergenliğe ulaştığında beyin gelişimini tamamlanmış olur		1	2	3	4	5	6	7	8	9	10	11	12
M27	İnsan beyninin normal gelişimi, beyin hücrelerinin doğumunu ve ölümünü içerir		1	2	3	4	5	6	7	8	9	10	11	12
M28	Uyuduğumuzda beynimiz çalışmayı durdurur		1	2	3	4	5	6	7	8	9	10	11	12
M29	Ortalama olarak, erkeklerin beyinleri kadınlardan daha büyüktür		1	2	3	4	5	6	7	8	9	10	11	12
M30	İnsanlar yaşamları süresince sahip olacakları tüm nöronlarla doğarlar		1	2	3	4	5	6	7	8	9	10	11	12



## Appendix 3. Neuromyths items learning process and intelligence characteristics: Arabic version.

البنود	الرقم	لكل بند يرجى كتابة المصير الذي أثر على اجاباتكم													
		صحيح	غير صحيح	التعليم الذي اتقاؤه بالجامعة	برامج التطوير المهني	خبرتي المهنية	وسائل التواصل الاجتماعي	مواقع الكترونية	من الزملاء	من نشرات العلمية	منشورات أخرى - كتب مجلات	من الأفلام	من التلفزيون	من الإعلانات	يبدو لي منطقياً
M1	يتعلم الطلاب بشكل أفضل عندما يتلقون المعلومات بأسلوب التعلم المفضل لديهم )على سبيل المثال، السمعي أو البصري أو الحركي(			1	2	3	4	5	6	7	8	9	10	11	12
M2	يجب أن يؤخذ في الاعتبار ملف الذكاء السائد للمتعلمين )على سبيل المثال، الذكاء المنطقي الرياضي، اللفظي، المكاني (في التدريس			1	2	3	4	5	6	7	8	9	10	11	12
M3	في عملية التعلم، يربط الدماغ المعلومات الجديدة بالمعرفة السابقة			1	2	3	4	5	6	7	8	9	10	11	12
M4	تعمل أجزاء مختلفة من الدماغ بشكل منفصل عن بعضها البعض أثناء عملية التعلم			1	2	3	4	5	6	7	8	9	10	11	12
M5	يحدث التعلم من خلال التغيرات في الاتصالات المتشابكة بين الخلايا العصبية في الدماغ			1	2	3	4	5	6	7	8	9	10	11	12
M6	التعلم مهارة معرفية بحتة وليست مهارة عاطفية			1	2	3	4	5	6	7	8	9	10	11	12
M7	يتم التعلم بشكل مستقل عن مسارات تعلم الأفراد			1	2	3	4	5	6	7	8	9	10	11	12
M8	يمكن لبعض العمليات العقلية )الخبرة والتعلم( المتكررة على مدى فترة طويلة أن تعدل بنية وعمل مناطق معينة من الدماغ			1	2	3	4	5	6	7	8	9	10	11	12
M9	يمكن للإنسان أن يتعلم معلومات جديدة أثناء النوم			1	2	3	4	5	6	7	8	9	10	11	12
M10	إن العقل البشري متوافق بشكل خاص مع المهام المتعددة			1	2	3	4	5	6	7	8	9	10	11	12
M11	بعض الناس يستخدمون دماغهم الأيمن والبعض الآخر يميلون إلى دماغهم الأيسر، مما يساعد في تفسير الاختلافات في الطريقة التي نتعلم بها			1	2	3	4	5	6	7	8	9	10	11	12
M12	يتعلم الأفراد بشكل أفضل عندما يتم تقديم محتوى الدورة التدريبية في جلسات أو وحدات قصيرة			1	2	3	4	5	6	7	8	9	10	11	12
M13	هناك فترات محددة من الطفولة لا يمكن بعدها تعلم أشياء معينة			1	2	3	4	5	6	7	8	9	10	11	12
M14	الحفظ ليس له أي تأثير على عملية التعلم			1	2	3	4	5	6	7	8	9	10	11	12
M15	البيئات التي توفر قدرًا أكبر من المحفزات تعمل على تحسين وظائف المخ لدى الأطفال في مرحلة ما قبل المدرسة			1	2	3	4	5	6	7	8	9	10	11	12
M16	القدرات العقلية وراثية ولا يمكن تغييرها بالبيئة أو الخبرة			1	2	3	4	5	6	7	8	9	10	11	12
M17	الاستماع إلى الموسيقى الكلاسيكية يساعد على تحسين القدرات العقلية			1	2	3	4	5	6	7	8	9	10	11	12
M18	عندما تتضرر منطقة واحدة من الدماغ، يمكن لمناطق أخرى أن تتولى وظيفتها			1	2	3	4	5	6	7	8	9	10	11	12
M19	يمكن لجلسات قصيرة من تمارين التنسيق أن تحسن وظائف المخ على سبيل المثال، لمس الكاحل الأيمن باليد اليسرى والعكس(صحيح)			1	2	3	4	5	6	7	8	9	10	11	12
M20	نحن نستخدم حوالي 10 %فقط من قدرات دماغنا			1	2	3	4	5	6	7	8	9	10	11	12
M21	الأشخاص الذين لديهم أدمغة أكبر هم أكثر ذكاءً			1	2	3	4	5	6	7	8	9	10	11	12
M22	يستمر إنتاج اتصالات جديدة في الدماغ طوال الحياة			1	2	3	4	5	6	7	8	9	10	11	12
M23	تم تصميم أدمغة الذكور والإناث لأنواع مختلفة من المهارات			1	2	3	4	5	6	7	8	9	10	11	12
M24	الدماغ ينشط 24 ساعة في اليوم			1	2	3	4	5	6	7	8	9	10	11	12
M25	المكملات الغذائية مثل أوميغا 3 وأوميغا 6 لها تأثير إيجابي على الأداء المعرفي			1	2	3	4	5	6	7	8	9	10	11	12
M26	يكتمل نمو الدماغ عندما يصل الأفراد إلى نهاية سن البلوغ			1	2	3	4	5	6	7	8	9	10	11	12
M27	التطور الطبيعي للدماغ البشري ينطوي على ولادة وموت خلايا الدماغ			1	2	3	4	5	6	7	8	9	10	11	12
M28	عندما ننام، الدماغ لا يعمل			1	2	3	4	5	6	7	8	9	10	11	12
M29	في المتوسط، أدمغة الرجال أكبر من أدمغة النساء			1	2	3	4	5	6	7	8	9	10	11	12
M30	يولد البشر بكل الخلايا العصبية التي سيمتلكونها طوال حياتهم			1	2	3	4	5	6	7	8	9	10	11	12

## Appendix 4. Neuromyths items learning process and intelligence characteristics: Chinese version.

編號	題目	正確	不正確	根據每個題項，請勾選出您的回答受到了下列哪些因素的影響（可複選）：											
				1. 大學研究所教育	2. 專業發展課程如繼續教育課程或其它增	3. 個人專業經驗	4. 社群媒體(Twitter (X), Instagram,	5. 網路資訊	6. 同事或親友	7. 學術刊物	8. 其他刊物如書籍、期刊、雜誌	9. 電影	10. 電視節目	11. 廣告	12. 對我來說很合理
M1	當人們以其擅長或優勢的學習方法（如視覺、聽覺、動作學習）獲得資訊時，能夠學習得更好。			1	2	3	4	5	6	7	8	9	10	11	12
M2	個人的優勢智力（如：數學、語言、空間智力）必須納入教學時的考量。			1	2	3	4	5	6	7	8	9	10	11	12
M3	在學習歷程中，個人的心智運作會連結新資訊與既有知識。			1	2	3	4	5	6	7	8	9	10	11	12
M4	大腦的各個部分在學習歷程中彼此獨立工作。			1	2	3	4	5	6	7	8	9	10	11	12
M5	學習是透過大腦神經元之間的神經突觸連結變化發生的。			1	2	3	4	5	6	7	8	9	10	11	12
M6	學習是一種純粹的認知技能，而不是情緒技能			1	2	3	4	5	6	7	8	9	10	11	12
M7	學習的發生與個人的學習背景無關。			1	2	3	4	5	6	7	8	9	10	11	12
M8	一些長時間、重複的心智運作（如經驗、學習歷程）能夠改變某些大腦的結構與功能。			1	2	3	4	5	6	7	8	9	10	11	12
M9	人類能夠在睡著的情況下學習新知。			1	2	3	4	5	6	7	8	9	10	11	12
M10	人類是良好的多工處理者。			1	2	3	4	5	6	7	8	9	10	11	12
M11	有些人有更多的右腦特質，有些人有更多的左腦特質；這樣的分類方式有助於我們解釋人們在學習方式上的差異。			1	2	3	4	5	6	7	8	9	10	11	12
M12	當學習內容以簡短的單元方式呈現時，人們能夠學習得更好。			1	2	3	4	5	6	7	8	9	10	11	12
M13	童年時期存在特定的學習階段，若錯過就無法再學習。			1	2	3	4	5	6	7	8	9	10	11	12
M14	記憶對學習歷程沒有影響。			1	2	3	4	5	6	7	8	9	10	11	12
M15	提供大量刺激的環境能夠改善學齡前兒童大腦的功能。			1	2	3	4	5	6	7	8	9	10	11	12
M16	心智能力是遺傳的，無法透過環境或學習經驗改變。			1	2	3	4	5	6	7	8	9	10	11	12
M17	聽古典樂能夠改善心智能力。			1	2	3	4	5	6	7	8	9	10	11	12
M18	當大腦某個區域受損時，大腦其他部分可以接管它的功能。			1	2	3	4	5	6	7	8	9	10	11	12
M19	短期的協調性運動能夠改善大腦的功能。			1	2	3	4	5	6	7	8	9	10	11	12
M20	我們只使用了10%的大腦。			1	2	3	4	5	6	7	8	9	10	11	12
M21	大腦體積越大的人越聰明。			1	2	3	4	5	6	7	8	9	10	11	12
M22	大腦產生新的神經連結的現象會在整個生命歷程中持續進行。			1	2	3	4	5	6	7	8	9	10	11	12
M23	男性和女性的大腦是為了不同類型技能而設計			1	2	3	4	5	6	7	8	9	10	11	12
M24	大腦在一天24小時都保持活動的狀態。			1	2	3	4	5	6	7	8	9	10	11	12
M25	營養補充品如：omega-3和omega-6對學業成就有正向效果。			1	2	3	4	5	6	7	8	9	10	11	12
M26	大腦在青春期結束時就已發展完成。			1	2	3	4	5	6	7	8	9	10	11	12
M27	正常的大腦發展涉及腦細胞的誕生與死亡。			1	2	3	4	5	6	7	8	9	10	11	12
M28	當我們睡覺時，大腦會停止運作。			1	2	3	4	5	6	7	8	9	10	11	12
M29	平均而言，男性比女性擁有較大的大腦。			1	2	3	4	5	6	7	8	9	10	11	12
M30	人們在出生時就擁有他／她一生中所需的所有大腦神經元。			1	2	3	4	5	6	7	8	9	10	11	12

**Appendix 5.** Neuromyths items learning process and intelligence characteristics: French version.

No	Items	Correct	Incorrect	Pour chaque item, veuillez indiquer les sources qui ont influencé vos réponses (Vous pouvez cocher plus d'une réponse).											
				1. D'enseignements reçus à l'université	2. De programmes de développement professionnel	3. De mon expérience professionnelle	4. De médias sociaux (Twitter (X), Instagram, LinkedIn, YouTube, etc.)	5. De sites web	6. De mes collègues	7. De publications scientifiques	8. D'autres publications telles que des livres, des revues, des magazines non scolaires	9. De films	10. D'émissions de télévision	11. De publicités	12. Cela me paraît logique
M1	Les élèves apprennent mieux lorsqu'ils reçoivent l'information dans leur style d'apprentissage préféré (p. ex., auditif, visuel ou kinesthésique)			1	2	3	4	5	6	7	8	9	10	11	12
M2	Le profil d'intelligence prédominant des apprenants (p. ex., logico-mathématique, verbale, spatiale) doit être pris en compte dans l'enseignement			1	2	3	4	5	6	7	8	9	10	11	12
M3	Dans le processus d'apprentissage, le cerveau associe les nouvelles informations aux connaissances antérieures			1	2	3	4	5	6	7	8	9	10	11	12
M4	Les différentes parties du cerveau fonctionnent séparément les unes des autres au cours du processus d'apprentissage			1	2	3	4	5	6	7	8	9	10	11	12
M5	L'apprentissage se produit par des changements dans les connexions synaptiques entre les neurones du cerveau			1	2	3	4	5	6	7	8	9	10	11	12
M6	L'apprentissage est une compétence purement cognitive et non émotionnelle			1	2	3	4	5	6	7	8	9	10	11	12
M7	L'apprentissage se fait indépendamment des parcours d'apprentissage des individus			1	2	3	4	5	6	7	8	9	10	11	12
M8	Certains processus mentaux (expérience, apprentissage) répétés sur une longue période peuvent modifier la structure et le fonctionnement de certaines zones du cerveau			1	2	3	4	5	6	7	8	9	10	11	12
M9	Les personnes peuvent apprendre de nouvelles informations en dormant			1	2	3	4	5	6	7	8	9	10	11	12
M10	Le cerveau des humains est particulièrement compatible avec le multi-tâches			1	2	3	4	5	6	7	8	9	10	11	12
M11	Certaines personnes sont plutôt « cerveau droit » et d'autres plutôt « cerveau gauche », ce qui contribue à expliquer les différences dans la manière dont on apprend			1	2	3	4	5	6	7	8	9	10	11	12
M12	Les individus apprennent mieux lorsque le contenu du cours est présenté sous forme de courtes sessions ou de modules			1	2	3	4	5	6	7	8	9	10	11	12
M13	Il existe des périodes spécifiques de l'enfance après lesquelles certaines choses ne peuvent plus être apprises			1	2	3	4	5	6	7	8	9	10	11	12
M14	La mémorisation n'a aucun impact sur le processus d'apprentissage			1	2	3	4	5	6	7	8	9	10	11	12
M15	Les environnements qui offrent une plus grande quantité de stimuli améliorent le fonctionnement du cerveau des enfants d'âge préscolaire			1	2	3	4	5	6	7	8	9	10	11	12
M16	Les capacités mentales sont héréditaires et ne peuvent être modifiées par l'environnement ou l'expérience.			1	2	3	4	5	6	7	8	9	10	11	12
M17	Écouter de la musique classique permet d'améliorer les capacités mentales			1	2	3	4	5	6	7	8	9	10	11	12
M18	Quand une région du cerveau est endommagée, d'autres peuvent prendre en charge sa fonction			1	2	3	4	5	6	7	8	9	10	11	12
M19	De courtes séances d'exercices de coordination peuvent améliorer le fonctionnement du cerveau (par exemple, toucher la cheville droite avec la main gauche et vice versa)			1	2	3	4	5	6	7	8	9	10	11	12
M20	Nous n'utilisons environ que 10 % de notre cerveau			1	2	3	4	5	6	7	8	9	10	11	12
M21	Les personnes ayant un cerveau plus gros sont plus intelligentes			1	2	3	4	5	6	7	8	9	10	11	12
M22	La production de nouvelles connexions dans le cerveau se poursuit tout au long de la vie			1	2	3	4	5	6	7	8	9	10	11	12
M23	Les cerveaux masculins et féminins sont conçus pour différents types de compétences			1	2	3	4	5	6	7	8	9	10	11	12
M24	Le cerveau est actif 24 heures sur 24			1	2	3	4	5	6	7	8	9	10	11	12
M25	Les suppléments tels que les Oméga-3 et les Oméga-6 ont un effet positif sur les performances cognitives			1	2	3	4	5	6	7	8	9	10	11	12
M26	Le développement du cerveau est achevé lorsque les individus atteignent la fin de la puberté			1	2	3	4	5	6	7	8	9	10	11	12
M27	Le développement normal du cerveau humain implique la naissance et la disparition de cellules cérébrales			1	2	3	4	5	6	7	8	9	10	11	12
M28	Quand on dort, le cerveau ne fonctionne pas			1	2	3	4	5	6	7	8	9	10	11	12
M29	En moyenne, le cerveau des hommes est plus gros que celui des femmes			1	2	3	4	5	6	7	8	9	10	11	12
M30	L'humain naît avec tous les neurones dont il disposera tout au long de sa vie			1	2	3	4	5	6	7	8	9	10	11	12

## Appendix 6. Neuromyths items learning process and intelligence characteristics: Greek version.

Αρ.	Προτάσεις	Σωστό	Λάθος	Για κάθε στοιχείο, αναφέρετε ποιες πηγές επηρέασαν τις απαντήσεις σας (Μπορείτε να ελέγξετε περισσότερες από μία απαντήσεις)											
				1. Προπτυχιακή/πτυχιακή εκπαίδευση	2. Προγράμματα επαγγελματικής ανάπτυξης	3. Η επαγγελματική μου εμπειρία	4. Μία κοινωνική δικτύωση (Twitter (X), Instagram, LinkedIn, YouTube etc.)	5. Διαδικτυακοί τόποι	6. Συνάδελφοι - Φίλοι	7. Ακαδημαϊκές δημοσιεύσεις	8. Άλλες εκδόσεις, όπως βιβλία, εκπαιδευτικά ή διδακτικά τευχικά	9. Κριματογράφος	10. Τηλεοπτικά προγράμματα	11. Διαφημίσεις	12. Διάθεση
M1	Τα άτομα μαθαίνουν καλύτερα όταν λαμβάνουν πληροφορίες με βάση τα διαφορετικά μαθησιακά τους προφίλ (πχ: οπτικό, ακουστικό, κιναισθητικό, κλπ.).			1	2	3	4	5	6	7	8	9	10	11	12
M2	Το κυρίαρχο προφίλ νοημοσύνης των μαθητών (πχ: λογικομαθηματικό, γλωσσικό, χωροταξικό) πρέπει να λαμβάνεται υπόψη στη διδασκαλία.			1	2	3	4	5	6	7	8	9	10	11	12
M3	Κατά τη διαδικασία της μάθησης, το μυαλό συσχετίζει νέες πληροφορίες με προηγούμενες γνώσεις.			1	2	3	4	5	6	7	8	9	10	11	12
M4	Διαφορετικά μέρη του εγκεφάλου λειτουργούν ανεξάρτητα κατά τη διάρκεια της μαθησιακής διαδικασίας.			1	2	3	4	5	6	7	8	9	10	11	12
M5	Η μάθηση προκύπτει από τις αλλαγές μεταξύ των νευρικών συνάψεων στον εγκέφαλο.			1	2	3	4	5	6	7	8	9	10	11	12
M6	Η μάθηση είναι μια καθαρά γνωστική δεξιότητα, όχι συναισθηματική.			1	2	3	4	5	6	7	8	9	10	11	12
M7	Η μάθηση είναι ανεξάρτητη από το γνωστικό υπόβαθρο των ατόμων.			1	2	3	4	5	6	7	8	9	10	11	12
M8	Μερικές νοητικές διαδικασίες (εμπειρία, μάθηση) που επαναλαμβάνονται για μεγάλο χρονικό διάστημα μπορούν να αλλάξουν τη δομή και τη λειτουργία ορισμένων περιοχών του εγκεφάλου.			1	2	3	4	5	6	7	8	9	10	11	12
M9	Τα άτομα μπορούν να μάθουν νέες πληροφορίες ακόμα και όταν βρίσκονται σε κατάσταση ύπνου.			1	2	3	4	5	6	7	8	9	10	11	12
M10	Οι άνθρωποι είναι καλοί στο να κάνουν πολλά πράγματα ταυτόχρονα.			1	2	3	4	5	6	7	8	9	10	11	12
M11	Το γεγονός ότι μερικοί άνθρωποι έχουν πιο ανεπτυγμένο το δεξιό ημισφαίριο του εγκεφάλου ενώ άλλοι το αριστερό, βοηθά στο να εξηγήσουμε τις διαφορές στο πώς μαθαίνουμε.			1	2	3	4	5	6	7	8	9	10	11	12
M12	Τα άτομα μαθαίνουν καλύτερα όταν το περιεχόμενο του μαθήματος παρουσιάζεται σε σύντομες περιόδους ή ενότητες.			1	2	3	4	5	6	7	8	9	10	11	12
M13	Υπάρχουν συγκεκριμένες περίοδοι στην παιδική ηλικία μετά τις οποίες ορισμένα πράγματα δεν μπορούν πλέον να αποτελέσουν αντικείμενο μάθησης.			1	2	3	4	5	6	7	8	9	10	11	12
M14	Η απομνημόνευση δεν έχει καμία επίδραση στη διαδικασία μάθησης.			1	2	3	4	5	6	7	8	9	10	11	12
M15	Περιβάλλοντα που προσφέρουν περισσότερα ερεθίσματα βελτιώνουν τον εγκέφαλο των παιδιών προσχολικής ηλικίας.			1	2	3	4	5	6	7	8	9	10	11	12
M16	Η διανοητική ικανότητα είναι κληρονομική και δεν μπορεί να αλλάξει από το περιβάλλον ή την εμπειρία.			1	2	3	4	5	6	7	8	9	10	11	12
M17	Η ακρόαση κλασικής μουσικής βελτιώνει την διανοητική ικανότητα.			1	2	3	4	5	6	7	8	9	10	11	12
M18	Όταν ένα μέρος του εγκεφάλου έχει υποστεί βλάβη, άλλα μέρη μπορούν να αναλάβουν τη λειτουργία του.			1	2	3	4	5	6	7	8	9	10	11	12
M19	Σύντομες περίοδοι ασκήσεων συντονισμού μπορούν να βελτιώσουν τη λειτουργία του εγκεφάλου (για παράδειγμα, το να αγγίζεις το δεξιό αστράγαλό με το αριστερό χέρι και το αντίστροφο).			1	2	3	4	5	6	7	8	9	10	11	12
M20	Χρησιμοποιούμε μόνο το 10% του εγκεφάλου μας.			1	2	3	4	5	6	7	8	9	10	11	12
M21	Τα άτομα με μεγαλύτερους εγκεφάλους είναι εξυπνότερα.			1	2	3	4	5	6	7	8	9	10	11	12
M22	Ο εγκέφαλος συνεχίζει να δημιουργεί νέες συνδέσεις κατά τη διάρκεια της ζωής του ατόμου.			1	2	3	4	5	6	7	8	9	10	11	12
M23	Οι εγκέφαλοι των ανδρών και των γυναικών είναι σχεδιασμένοι για διαφορετικούς τύπους δεξιοτήτων.			1	2	3	4	5	6	7	8	9	10	11	12
M24	Ο εγκέφαλος παραμένει ενεργός 24 ώρες το εικοσιτετράωρο.			1	2	3	4	5	6	7	8	9	10	11	12
M25	Συμπληρώματα διατροφής όπως Ωμέγα-3 και Ωμέγα-6 έχουν θετική επίδραση στην ακαδημαϊκή επιτυχία.			1	2	3	4	5	6	7	8	9	10	11	12
M26	Η ανάπτυξη του εγκεφάλου έχει ολοκληρωθεί μέχρι το τέλος της εφηβείας.			1	2	3	4	5	6	7	8	9	10	11	12
M27	Η φυσιολογική ανάπτυξη του ανθρώπινου εγκεφάλου περιλαμβάνει τη γέννηση και το θάνατο των εγκεφαλικών κυττάρων.			1	2	3	4	5	6	7	8	9	10	11	12
M28	Ο εγκέφαλος «κλείνει» κατά τη διάρκεια του ύπνου.			1	2	3	4	5	6	7	8	9	10	11	12
M29	Κατά μέσο όρο, οι άνδρες έχουν μεγαλύτερο εγκέφαλο από τις γυναίκες.			1	2	3	4	5	6	7	8	9	10	11	12
M30	Οι άνθρωποι γεννιούνται με όλους τους νευρώνες που θα έχουν κατά τη διάρκεια της ζωής τους.			1	2	3	4	5	6	7	8	9	10	11	12

## Appendix 7. Neuromyths items learning process and intelligence characteristics: Kazakh version.

Ар.	Протоқас	Σοστό	Απόος	Өрбір сұраққа берген жауабыңызға қандай дереккөз әсер еткенін көрсетіңіз (бірнеше дереккөзді таңдауыңызға болады)											
				1. Бакалавриат/магистратура білімінен	2. Кәсіби даму бағдарламаларынан	3. Кәсіби тәжірибемнен	4. Әлеуметтік желілер (Мысалы: Twitter (X), Instagram, LinkedIn, YouTube)	5. Веб-сайттардан	6. Әріптестерімнен немесе достарымнан	7. Академиялық басшылардан	8. Кітаптар, журналдар, танымал журналдар сияқты басқа басшылардан	9. Фильмдерден	10. Теледидар бағдарламаларынан	11. Жарнамалардан	12. Мен бұл тұжырыммен келісемін
M1	Адамдар ақпаратты өздерінің басымдығы жоғары қабылдау дағдыларына сай жақсырақ меңгереді. (мысалы: аудиал, визуал, кинестетикалық) (			1	2	3	4	5	6	7	8	9	10	11	12
M2	Оқу (үйрену) кезінде оқушылардың басым интеллектуал дағдысы (мысалы: математикалық, сөздік, кеңістіктік) ескерілуі керек			1	2	3	4	5	6	7	8	9	10	11	12
M3	Оқу (үйрену) үдерісінде ақыл-ой жаңа ақпаратты бұрынғы білімімен байланыстырады			1	2	3	4	5	6	7	8	9	10	11	12
M4	Оқу (үйрену) процесінде ми бөлімдері бір-бірінен бөлек-бөлек жұмыс істейді			1	2	3	4	5	6	7	8	9	10	11	12
M5	Оқу (үйрену) мидағы нейрондар арасындағы синапс байланыстардың өзгеруі арқылы жүреді			1	2	3	4	5	6	7	8	9	10	11	12
M6	Оқу (үйрену) эмоциялық емес, таза когнитивті дағды			1	2	3	4	5	6	7	8	9	10	11	12
M7	Оқу (үйрену) жеке тұлғалардың білім деңгейіне қарамастан жүзеге асады			1	2	3	4	5	6	7	8	9	10	11	12
M8	Ұзақ уақыт бойы қайталанатын кейбір ақыл-ой үдерісі (тәжірибе, оқу) мидың кейбір аймақтарының құрылымы мен қызметін өзгертуі мүмкін			1	2	3	4	5	6	7	8	9	10	11	12
M9	Адамдар ұйықтап жатқанда жаңа ақпаратты меңгере алады			1	2	3	4	5	6	7	8	9	10	11	12
M10	Адамдар көп тапсырманы бір мезетте жақсы орындай алады			1	2	3	4	5	6	7	8	9	10	11	12
M11	Кейбір адамдардың «оң жақ ми сыңары», ал басқаларының «сол жақ ми сыңары» жақсы жұмыс істейді. Бұл біздің оқу (үйрену) жолындағы айырмашылықтарымызды түсіндіруге көмектеседі			1	2	3	4	5	6	7	8	9	10	11	12
M12	Адамдар мазмұны қысқа сессиялардан немесе модульдерден тұратын курстарды жақсырақ меңгереді (үйренеді)			1	2	3	4	5	6	7	8	9	10	11	12
M13	Балалық шақта белгілі бір ерекше кезеңдер болады. Кейбір нәрселерді сол аралықта ғана үйренуге болады			1	2	3	4	5	6	7	8	9	10	11	12
M14	Есте сақтау қабілеті оқу (үйрену) процесіне әсер етпейді			1	2	3	4	5	6	7	8	9	10	11	12
M15	Көбірек ынталандыратын орталар мектеп жасына дейінгі балалардың миын жақсартады			1	2	3	4	5	6	7	8	9	10	11	12
M16	Ақыл-ой қабілеті тұқым қуалайды және оның қоршаған ортасы немесе өмірлік тәжірибесі өзгерте алмайды			1	2	3	4	5	6	7	8	9	10	11	12
M17	Классикалық музыканы тыңдау ақыл-ой қабілетін жақсартады			1	2	3	4	5	6	7	8	9	10	11	12
M18	Мидың бір аймағы зақымдалғанда, оның функциясын мидың басқа аймақтары орындауға кіріседі			1	2	3	4	5	6	7	8	9	10	11	12
M19	Қысқа мерзімді үйлестіру жаттығулары ми қызметін жақсартады			1	2	3	4	5	6	7	8	9	10	11	12
M20	Біз миымыздың тек 10%-ын пайдаланамыз			1	2	3	4	5	6	7	8	9	10	11	12
M21	Миы үлкен адамдар ақылдырақ			1	2	3	4	5	6	7	8	9	10	11	12
M22	Мидағы жаңа байланыстардың қалыптасуы өмір бойы жалғасады			1	2	3	4	5	6	7	8	9	10	11	12
M23	Ерлер мен әйелдердің миы әртүрлі дағдыларға арналған			1	2	3	4	5	6	7	8	9	10	11	12
M24	Ми тәулігіне 24 сағат белсенді			1	2	3	4	5	6	7	8	9	10	11	12
M25	Омега-3 және омега-6 сияқты қоспалар оқу жетістіктеріне оң әсер етеді			1	2	3	4	5	6	7	8	9	10	11	12
M26	Балалардың жыныстық жетілуі аяқталғанда, мидың дамуы да соңына жетеді			1	2	3	4	5	6	7	8	9	10	11	12
M27	Адам миының қалыпты дамуы ми жасушаларының тууы мен өлуінен тұрады			1	2	3	4	5	6	7	8	9	10	11	12
M28	Біз ұйықтап жатқанда, ми өз жұмысын тоқтатады			1	2	3	4	5	6	7	8	9	10	11	12
M29	Орташа алғанда, ерлердің миы әйелдерге қарағанда үлкенірек			1	2	3	4	5	6	7	8	9	10	11	12
M30	Адам баласы оның өмірінде болатын барлық нейрондарды толық иемдене туады			1	2	3	4	5	6	7	8	9	10	11	12



**Appendix 8.** *Neuromyths items learning process and intelligence characteristics: Malay version.*

		Betul	tak betul	Untuk setiap item, sila tunjukkan sumber yang paling mempengaruhi jawapan and (anda boleh memilih lebih daripada satu jawapan)											
				1. Pendidikan sarjana muda/ Pendidikan	2. Program perkemabangan profesional	3. Pengalaman profesional	4. Media sosial	5. laman web	6. Rakan sekerja atau kawan	7. Penerbitan akademik	8. Penerbitan lain seperti buku, journal, majalah popular	9. Wayang	10. televisyen	11. Iklan	12. Intuisi saya
M1	Individu belajar dengan lebih baik bila mereka menerima maklumat yang serasi dengan gaya pembelajaran dominan mereka (contohnya: visual, auditori, kinestetik dll)			1	2	3	4	5	6	7	8	9	10	11	12
M2	Profil kecerdasan pelajar yang dominan (contohnya: matematik, verbal, spatial) mesti dipertimbangkan semasa pengajaran.			1	2	3	4	5	6	7	8	9	10	11	12
M3	Dalam proses pembelajaran, minda kita akan mengaitkan maklumat baru dengan pengetahuan yang terdahulu.			1	2	3	4	5	6	7	8	9	10	11	12
M4	Bahagian otak yang berlainan akan beroperasi secara bersendirian semasa proses pembelajaran.			1	2	3	4	5	6	7	8	9	10	11	12
M5	Proses pembelajaran berlaku melalui perubahan dalam sambungan sinaptik antara neuron dalam otak.			1	2	3	4	5	6	7	8	9	10	11	12
M6	Pembelajaran hanya melibatkan kemahiran kognisi dan bukan emosi.			1	2	3	4	5	6	7	8	9	10	11	12
M7	Pembelajaran berlaku secara bebas dari (tidak dipengaruhi oleh) latar belakang individu.			1	2	3	4	5	6	7	8	9	10	11	12
M8	Semua proses mental (pengalaman, pembelajaran) yang berulang dalam jangka masa yang panjang akan mengubah struktur dan fungsi bahagian tertentu otak.			1	2	3	4	5	6	7	8	9	10	11	12
M9	Individu akan belajar maklumat baru walaupun semasa tidur.			1	2	3	4	5	6	7	8	9	10	11	12
M10	Manusia boleh melakukan pelbagai tugas dalam masa yang sama secara baik.			1	2	3	4	5	6	7	8	9	10	11	12
M11	Hakikat bahawa orang tertentu lebih cenderung kepada 'otak kanan' dan yang lain 'otak kiri' membantu menerangkan perbezaan dalam cara kita belajar.			1	2	3	4	5	6	7	8	9	10	11	12
M12	Individu belajar dengan lebih baik bila kandungan kursus dipersembahkan dalam sesi pendidikan atau secara bermodul.			1	2	3	4	5	6	7	8	9	10	11	12
M13	Terdapat perkara tentu yang boleh dipelajari semasa peringkat kanak-kanak yang tidak dapat dipelajari lagi dalam umur lain			1	2	3	4	5	6	7	8	9	10	11	12
M14	Penghafalan tiada impak ke atas proses pembelajaran.			1	2	3	4	5	6	7	8	9	10	11	12
M15	Persekitaran yang mampu membekalkan rangsangan yang banyak boleh meningkatkan fungsi otak di kalangan kanak-kanak pra sekolah.			1	2	3	4	5	6	7	8	9	10	11	12
M16	Kapasiti mental adalah secara keturunan dan tidak boleh diubah oleh persekitaran atau pengalaman.			1	2	3	4	5	6	7	8	9	10	11	12
M17	Mendengar music klasikal meningkatkan kapasiti mental			1	2	3	4	5	6	7	8	9	10	11	12
M18	Bila sebahagian daripada otak sudah rosak, bahagian lain boleh mengambil alih fungsinya			1	2	3	4	5	6	7	8	9	10	11	12
M19	Latihan koordinasi jangka pendek boleh meningkatkan fungsi otak (contohnya, menyentuh buku lali kanan dengan tangan kiri dan sebaliknya)			1	2	3	4	5	6	7	8	9	10	11	12
M20	Kita cuma gunakan 10% otak kita.			1	2	3	4	5	6	7	8	9	10	11	12
M21	Individu dengan otak yang lebih besar adalah lebih pandai			1	2	3	4	5	6	7	8	9	10	11	12
M22	Otak kita terus menerbitkan sambungan baru sepanjang umur individu			1	2	3	4	5	6	7	8	9	10	11	12
M23	Otak lelaki dan perempuan direkabentuk untuk jenis kemahiran yang berlainan			1	2	3	4	5	6	7	8	9	10	11	12
M24	Otak tetap aktif selama 24 jam sehari			1	2	3	4	5	6	7	8	9	10	11	12
M25	Makanan tambahan seperti Omega-3 dan Omega-6 mempunyai kesan positif ke atas pencapaian akademik			1	2	3	4	5	6	7	8	9	10	11	12
M26	Perkembangan otak adalah lengkap bila kanak-kanak mencapai akhir akil baligh			1	2	3	4	5	6	7	8	9	10	11	12
M27	Perkembangan biasa otak manusia melibatkan kelahiran dan kematian sel otak			1	2	3	4	5	6	7	8	9	10	11	12
M28	Otak tutup fungsi semasa tidur			1	2	3	4	5	6	7	8	9	10	11	12
M29	Secara lazim, lelaki mempunyai otak yang lebih besar daripada perempuan			1	2	3	4	5	6	7	8	9	10	11	12
M30	Manusia dilahirkan dengan semua neuron yang mereka ada dalam jangka hayat mereka			1	2	3	4	5	6	7	8	9	10	11	12

## Video annotation tools for assessing psychomotor skills in nursing education: A scoping review

Greet Leysens<sup>1\*</sup>, Rani Claus<sup>2\*</sup>, Wim Van Petegem<sup>3</sup>, Nathalie Charlier<sup>1</sup>

<sup>1</sup>KU Leuven, Educational Master in Health Sciences, and Department of Pharmaceutical and Pharmacological Sciences Leuven, Belgium

<sup>2</sup>KU Leuven, Department of Public Health and Primary Care, Environment and Health, Leuven, Belgium

<sup>3</sup>KU Leuven, Engineering Technology Education Research (ETHER), Group T Leuven Campus, Belgium

### ARTICLE HISTORY

Received: June 4, 2025

Accepted: Aug. 29, 2025

### Keywords:

Video-based assessment,  
Psychomotor skill  
acquisition,  
Technology-enhanced  
learning,  
Nursing education,  
Video annotation.

**Abstract:** Video annotation tools can improve assessments by enabling visualization, repeated reviews, and structured feedback, thereby promoting deeper learning. This approach is particularly relevant in health professional education, where acquiring psychomotor skills is essential. No recent reviews have focused specifically on video annotation tools for assessing psychomotor skills in nursing education. This scoping review explores tools that facilitate feedback and assessment of psychomotor performances, with a focus on those customizable for nursing-specific psychomotor skills. Included studies are published in peer-reviewed journals and utilize software that allows for annotation of psychomotor performances, integration of feedback criteria and customization for nursing education. Studies employing video annotation tools solely for machine learning algorithms or for non-psychomotor performances are excluded. Literature searches were conducted using PubMed, Web of Science Core collection, CINAHL via EBSCO, Scopus, EuropePMC, CENTRAL via Cochrane Library, ERIC via OVID, FiS Bildung, SportDiscus via EBSCO, IEEE and ACM digital library. Additional searches included snowballing and Google. The initial screening identified 18 video annotation tools. Following a second screening phase, seven tools were excluded due to the absence of essential features, such as support for multiple camera angles or isochronic annotation. Ultimately, we highlight four customizable tools that are particularly relevant to nursing education. This scoping review provides a springboard for the development of a tailored video annotation tool that builds on existing software. The aim of such a tool will be to streamline feedback and assessment processes, enhance learning outcomes for nursing students, and provide nursing lecturers with an efficient, practice-aligned solution.

## 1. INTRODUCTION

In nursing education, a key focus is teaching psychomotor skills, such as venipuncture or injection. These psychomotor nursing skills (PNS) are movement-oriented performances of varying complexity (INACSL, 2011). Students first practice PNS in a skills lab using task trainers, before applying them in complex real-life settings. This controlled environment allows for performance assessment, which is essential for learning through targeted feedback (Miles,

\*CONTACT: Greet LEYSENS and Rani CLAUS ✉ [greet.leysens@kuleuven.be](mailto:greet.leysens@kuleuven.be); [rani.claus@kuleuven.be](mailto:rani.claus@kuleuven.be)

📍 Tervuursevest 101 – bus 1500, 3001 Heverlee, Belgium

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

2018). Traditionally, feedback is provided verbally or in writing, but memory bias can affect its accuracy. Video-based assessment can enhance feedback (Mayer, 2014) by providing visual footage on students' performances, supporting more precise, actionable guidance and deeper skill mastery. This provides opportunities, especially since Gen Z healthcare students prefer visual learning environments such as video (Shorey *et al.*, 2021). Bahula & Kay (2021) found that students preferred video-based over text-based feedback due to its detailed, clear, and rich quality. Epstein *et al.* (2020) demonstrated that smartphone video for practice-based learning and PNS assessment enhanced deeper learning and flexible feedback in a blended learning environment.

To deliver feedback to students, annotating video footage can be an efficient approach. Lam & Habil (2021b) describe video annotation as “a tool, a learning system, a Web 2.0 application, a platform, a device, a software, a program or simply an application of technology associated with a feature which enables individuals to annotate audio-visual content, either with textual or multimedia annotations. Most video annotation platforms are characterized by a feature for segmentation or clipping of segments of a video with comments that are synchronized with the video timeline”. Video annotation positively impacts learning and professional development by facilitating reflection, comprehension, critical thinking, and satisfaction (Evi-Colombo *et al.*, 2020; Lam & Habil, 2021b). To deliver targeted feedback, various forms of annotation can be distinguished. Rolf *et al.* (2014) identify isochronic annotations, which link content to a specific time, spatial annotations, which link content to a point in an area, and structured annotations, which add textual comments to the video. To be particularly effective in supporting the learning and assessment of PNS, a video annotation tool should meet several key criteria. First, it should support isochronic annotation, enabling feedback to be synchronized with specific time points in a skill demonstration, for example, to highlight the exact moment a critical error occurs during a procedural step. Second, spatial annotation capabilities are essential to direct the learner's attention to a specific area of the screen, such as circling incorrect hand placement during an injection. Third, tools should offer structured annotation features, including predefined categories or labels (e.g., sterility error, incorrect execution), to facilitate consistent and targeted feedback. Additionally, tools should be adaptable to the nursing education context, meaning they can be used flexibly by lecturers and students, operate within the privacy and usability constraints of healthcare training environments, and allow customization to align with specific criteria of a specific nursing skill. These features enhance learning by enabling specific and actionable feedback, fostering reflection, and supporting deliberate practice, all critical for the acquisition of PNS.

Evi-Colombo *et al.* (2020) reviewed seventeen video annotation tools (VATs), highlighting their technical and pedagogical affordances in different disciplines. Lam & Habil (2021a) identified 20 VATs for video-annotated peer feedback. Both reviews describe VideoANT, Media Annotation Tool and Collaborative Lecture Annotation System as the three most researched VATs. VATs vary in features, such as color coding, written or drawn-on annotation, tagging or labelling and exporting options (Evi-Colombo *et al.*, 2020). VATs have been researched in teacher training to improve reflections on teaching practices (Ardley & Hallare, 2020; Ardley & Johnson, 2019; McFadden *et al.*, 2014; Nagel & Engeness, 2021; Pérez-Torregrosa *et al.*, 2017; Sain, 2022), and in blended learning to annotate lecture recordings for reflective and active learning (Aubert *et al.*, 2014; Cassano & Di Blas, 2023; Douglas *et al.*, 2014; Rolf *et al.*, 2014). VATs in sports can provide insights for evaluating and enhancing physical performance (O'Donoghue & Holmes, 2014; Shih, 2018). Barriers to VAT implementation include technical issues, training needs, video camera type, workload, accessibility, storage, and cost (Ardley & Hallare, 2020; Frehner *et al.*, 2012; Hands *et al.*, 2009; Rich & Trip, 2011). In health professional education, Hands *et al.* (2009) used Dartfish to tag videos for learning and assessment, reducing student anxiety and enhancing learning. Frehner *et al.* (2012) developed and explored a VAT for reviewing self- and peer-recorded

nursing skill performances, positively impacting reflective practices. Both projects encountered technical issues related to video storage. Future research should focus on the impact of learning outcomes and on developing a financially feasible and accessible tool to assess skill performances in higher education.

Despite the growing use of video in nursing education, a consolidated overview of VATs suitable for providing structured feedback on PNS is lacking. Existing reviews focus on general educational or peer-feedback contexts, but do not examine the specific features required for effective feedback on complex, stepwise psychomotor performances in healthcare education. As a result, lecturers face uncertainty when selecting or implementing VATs that align with both pedagogical needs and practical constraints.

This scoping review addresses that gap by mapping the current landscape of VATs and identifying tools with the technical and functional capacity to support accurate, customizable, and context-appropriate feedback aligned with the specific steps of each PNS (Vermeulen, 2019). To address this gap, we conducted a scoping review guided by the following research questions (RQs):

RQ 1: Which VATs can generate isochronic, spatial, and structured annotations to provide accurate visual feedback on movement-oriented performances?

RQ 2: Which of these VATs are customizable for specific PNS to provide visual, clear, precise and actionable feedback to students in nursing education?

## 2. METHOD

This scoping review, development of the review protocol and summary of evidence is performed using the JBI Manual for Evidence Synthesis (Peters *et al.*, 2024) and is in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (Tricco *et al.*, 2018). The study protocol is registered in the Research Data Repository (KU Leuven) (Leysens *et al.*, 2024, December 20).

### 2.1. Inclusion and Exclusion Criteria

Records are eligible for inclusion if they meet the following criteria:

- Utilize software capable of annotating psychomotor performances.
- Demonstrate potential for customization to annotate the performance of PNS, specifically by embedding feedback criteria.
- Are published in peer-reviewed journals in Dutch, English, French or German.
- There are no restrictions on the publication date.

Records are excluded if they describe a VAT:

- For machine learning algorithms, such as automated image annotation.
- To annotate videos unrelated to psychomotor performances (e.g. focus on cognitive or affective learning domains, medical imaging)
- For real-time video analytics with wearable sensors or trackers
- That does not allow annotation or is not customizable.
- That was no longer available at the time of this review process.

The Population-Concept-Context-Type description with inclusion and exclusion criteria is shown in [Appendix I](#).

### 2.2. Search Strategy

To answer RQ 1 and RQ 2, a literature search was conducted in PubMed, Web of Science core collection, CINAHL via EBSCO, Scopus, EuropePMC, CENTRAL via Cochrane Library, ERIC via OVID, FiS Bildung, SportDiscus via EBSCO, IEEE and ACM digital library from inception to February 5, 2024. An updated search was performed on January 8, 2025, to capture studies published since the initial search. The interdisciplinary nature of the study required a

broad database search, targeting full-text primary studies, reviews, meta-analyses, text and opinion papers, guidelines, and conference proceedings. Search strings were developed in collaboration with the Biomedical Library, 2Bergen (Leuven, Belgium) and are detailed in [Appendix II](#). Snowballing and grey literature searches were also performed, including a Google search using the term ‘annotation tool.’ The language was restricted to Dutch, English, French or German. An additional PubMed search using VAT names supplemented findings for RQ 2. Studies discussing the strengths and weaknesses of the selected VATs are included.

### 2.3. Source of Evidence Screening and Selection

After removing duplicates, titles and abstracts from the original search (up to February 5, 2024) were screened for eligibility by two independent reviewers (GL and RC), using Rayyan with a blind filter (Ouzzani *et al.*, 2016). The abstract screening followed best practice guidelines for scoping reviews (Polanin *et al.*, 2019). Discrepancies were resolved through discussion. Uncertain cases advanced to full-text screening. Relevant papers were retrieved in full, and imported into EndNote<sup>TM</sup> 21.2 (Clarivate Analytics, PA, USA). GL and RC independently reviewed full texts. Articles without accessible full text were excluded. For the updated search (January 8, 2025), GL applied the same eligibility criteria.

### 2.4. Data Extraction

During full-text screening, relevant data were extracted into a standardized Excel spreadsheet. The following data items were charted for each included study: name of the tool, target group, availability, and application of the tool. The extraction form was developed based on the review questions. The data extraction was performed by both reviewers, and discrepancies were resolved through discussion.

### 2.5. Analysis and Presentation of Results

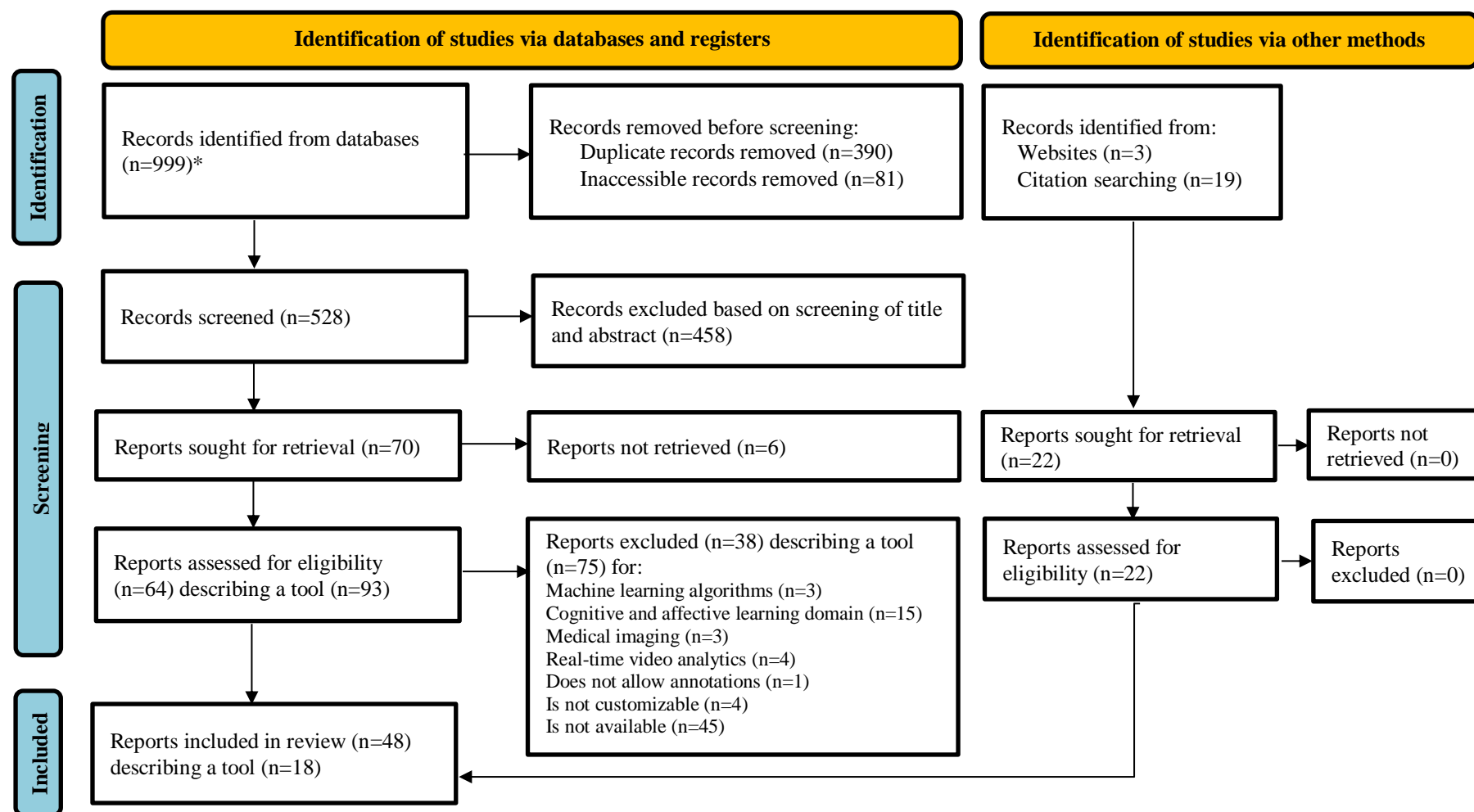
The data were synthesized using a descriptive, narrative approach aligned with the scoping review methodology. Extracted data were charted and grouped according to key characteristics of VATs relevant to RQ1 and RQ2. For RQ1, VATs capable of generating isochronic, spatial, and structured annotations for movement-oriented feedback were tabulated. Descriptive categories included VAT name, target group, language, open-source availability, footage used, annotation capabilities, adaptability for PNS, and source references. Where necessary, supplementary information was retrieved from official VAT websites. For RQ2, VATs requiring minimal adaptations to provide actionable feedback to nursing students were identified and summarized.

## 3. RESULTS

### 3.1. Search Results

The original search identified 999 records. After removing duplicates (n=390) and inaccessible records (n=81), 528 records remained. Title and abstract screening retained 70 reports for full-text review, with six reports excluded due to unavailability. An additional 22 records were identified through citation tracking and Google Search, leading to 86 reports for full-text screening and the assessment of 93 tools. Of these, 75 tools were excluded for different reasons, and led to the removal of 38 reports. Ultimately, 48 reports were included, mapping 18 VATs. The updated search did not identify any new records or tools. A detailed PRISMA flowchart (Page *et al.*, 2021) outlining the original search strategy is presented in [Figure 1](#).





**Figure 1.** Flowchart of the search strategy and results (dated 5/2/2024) according to PRISMA 2020 Statement (Page et al., 2021).

\* Search strategy per database is comprehensively reported and detailed in [Appendix II](#)

### 3.2. Evidence Sources for VATs

Ultimately, 48 articles and 18 VATs were eligible for data extraction to address RQ 1. The VATs Catapult and Reclipped were discovered via Google Search. Six review articles identified CoachNow, CaTool, Dartfish, ELAN, FEVA, GoReact, Hudl Sportscode, SiliconCOACH, Spark Motion Pro, Utilius Fairplay 5, VIA, and VIAN (Barris & Button, 2008; Evi-Colombo *et al.*, 2020; Lam & Habil, 2021a; Laughlin *et al.*, 2019; Liebermann *et al.*, 2002; Shrestha *et al.*, 2023). The remaining 42 articles covered VAT studies. One study per VAT was identified for Anvil 6.0, FEVA, and Observer XT. Two studies each referred to Hudl Sportscode and GoReact and three studies mentioned SiliconCOACH or LINC PLUS. Kinovea was used in 14 studies and Dartfish in 15 studies. Quality analysis of the research designs is beyond the scope of this review.

Table 1 presents a narrative summary of VAT characteristics, including target group, language, footage, open-source availability, type of annotation and references (RQ1). Key features such as split-screen functionality for multi-camera angles and annotation capabilities are highlighted. These inform the VAT suitability for customizing feedback on PNS performance (RQ2). All VATs are available in English. Utilius Fairplay 5 is also available in German, and LINC PLUS in Spanish. Eight VATs (Anvil 6.0, CaTool, ELAN, FEVA, Kinovea, LINC PLUS, VIA and VIAN) are open source and free of charge; the other VATs require a subscription or purchase. Eleven VATs (Catapult, CoachNow, Dartfish, FEVA, Hudl Sportscode, Kinovea, LINC PLUS, Observer XT, SiliconCOACH, Spark Motion Pro and Utilius) support split-screen. All VATs support structured annotation, which can be displayed next to or below the video. CaTool and VIAN lack isochronic annotation features, while ELAN, GoReact and Reclipped do not support spatial annotation. Seven VATs cannot be customized (Anvil 6.0, CaTool, ELAN, GoReact, Reclipped, VIA and VIAN).

### 3.3. Evidence Sources for VATs Eligible for PNS

In addressing RQ 2, we prioritized customizability, as each PNS involves specific sequential steps, making the tailoring of assessment criteria within the VAT crucial. We emphasized the need for clear and actionable feedback by evaluating the potential for annotating multiple video angles simultaneously and incorporating isochronic, spatial, and structured annotations. Table 1 shows that 11 VATs meet these criteria (Catapult, CoachNow, Dartfish, FEVA, Hudl Sportscode, Kinovea, LINC PLUS, Observer XT, SiliconCOACH, Spark Motion Pro, and Utilius Fairplay 5). These will be further investigated to answer RQ2. A PubMed search using each VAT's name yielded 233 articles. Ten studies discussed the strengths and weaknesses of the VAT, 44 were duplicates and 179 were irrelevant. Table 2 provides an overview of the specific features of these 11 VATs, which will be discussed further based on key questions for educators, as outlined by Rich and Trip (2011).

**Table 1.** Features of 18 VATs.

VAT	Target group	Language	Footage	Open source*	Annotation*			Customizable for PNS*	References
					isochronic	spatial	structured		
Anvil 6.0	Research: social sciences, human behavior, and digital technologies	English	Single	✓	✓	✓	✓	X	(Loukas <i>et al.</i> , 2020; Shrestha <i>et al.</i> , 2023)
Catapult	Sport performances	English	Multi	X	✓	✓	✓	✓	Google Search
CaTool	Academia collaborative annotation	English	Single	✓	X	✓	✓	X	(Lam & Habil, 2021a)
CoachNow	Sport performances	English	Multi	X	✓	✓	✓	✓	(Laughlin <i>et al.</i> , 2019)
Dartfish	Sport performances	English	Multi	X	✓	✓	✓	✓	(Abdelrasoul <i>et al.</i> , 2015; Andrews & Bressan, 2018; Barris & Button, 2008; Bobo <i>et al.</i> , 2012; Chiappedi <i>et al.</i> , 2012); (Earp <i>et al.</i> , 2016; Hands <i>et al.</i> , 2009; Judge <i>et al.</i> , 2008; Liebermann <i>et al.</i> , 2002; Maykut <i>et al.</i> , 2015; Myer <i>et al.</i> , 2012; Ong <i>et al.</i> , 2015; Post <i>et al.</i> , 2016; Rucci & Tomporowski, 2010; Schärer <i>et al.</i> , 2021; Ste-Marie <i>et al.</i> , 2016; Ste-Marie <i>et al.</i> , 2012; Walker <i>et al.</i> , 2020)
ELAN	Linguistic annotations	English	Single	✓	✓	X	✓	X	(Shrestha <i>et al.</i> , 2023)
FEVA	Computer and social sciences	English	Multi	✓	✓	✓	✓	✓	(Shrestha <i>et al.</i> , 2023)
Go-React	Education, skills-based learning	English	Single	X	✓	X	✓	X	(Ardley & Hallare, 2020; Ardley & Johnson, 2019; Evi-Colombo <i>et al.</i> , 2020; Lam & Habil, 2021a; Schulz & Gaudreault, 2023)
Hudl Sportscode	Sport performances	English	Multi	X	✓	✓	✓	✓	(Beseler <i>et al.</i> , 2024; Peeters <i>et al.</i> , 2019)
Kinovea	Sport performances	English	Multi	✓	✓	✓	✓	✓	(Amri-Dardari <i>et al.</i> , 2022; Amri-Dardari <i>et al.</i> , 2020; Cabarkapa <i>et al.</i> , 2021; Carzoli <i>et al.</i> , 2022; Dadashi <i>et al.</i> , 2013; Gonzalvo <i>et al.</i> , 2017; Ishac

									& Eager, 2021; Puklavec <i>et al.</i> , 2021; Raiola <i>et al.</i> , 2013; Souissi <i>et al.</i> , 2023; Souissi <i>et al.</i> , 2021; Tayech <i>et al.</i> , 2022; Yang <i>et al.</i> , 2022)
LINCE PLUS	Sport performances	English Spanish	Multi	✓	✓	✓	✓	✓	(Prieto-Lage <i>et al.</i> , 2020; Soto <i>et al.</i> , 2019; Soto-Fernández <i>et al.</i> , 2021)
Observer XT	Research: behavioral	English	Multi	X	✓	✓	✓	✓	(Dove & Astell, 2019)
Reclipped	Individual or team note taking; research; analysis and feedback	English	Single	X	✓	X	✓	X	Google Search
Silicon-COACH	Sport performances; sports retail; education; clinical practice	English	Multi	X	✓	✓	✓	✓	(Lago-Fuentes <i>et al.</i> , 2018; Liebermann <i>et al.</i> , 2002; McDonald <i>et al.</i> , 2011; Shultz <i>et al.</i> , 2013)
Spark Motion Pro	Sport performances; medical practitioners	English	Multi	X	✓	✓	✓	✓	(Laughlin <i>et al.</i> , 2019)
Utilius Fairplay 5	Sport performances	German English	Multi	X	✓	✓	✓	✓	(Barris & Button, 2008)
VIA	Academia; commercial	English	Single	✓	✓	✓	✓	X	(Shrestha <i>et al.</i> , 2023)
VIAN	Film analysis	English	Single	✓	X	✓	✓	X	(Shrestha <i>et al.</i> , 2023)

\* ✓ = yes; X = no

**Table 2.** Features of eligible VAT for PNS based on information from the VAT website and articles.

VAT	Cost*	Short Keys*	Live annotation*	Speed changing*	Collaboration*	Data management*	App*	Support*	Output	Extra PubMed
Catapult	✓	?	✓	?	✓	✓	✓	✓	Custom workbook views with preset filters; 2D pitch and graphs	0
CoachNow	✓	✓	X	✓	✓	✓	✓	✓	Annotated video with possibility to attach spreadsheets, documents, notes and comments	0
Dartfish	✓	✓	✓	✓	✓	✓	✓	✓	Video with written annotations and coded segments; export as CSV, image, video, and report	75
FEVA	X	✓	X	✓	X	✓	X	X	Export as video, image, and closed caption	0
Hudl Sportscode	✓	✓	✓	✓	✓	✓	✓	✓	Annotated clips; statistical breakdowns; visualized performance data; tailored detailed reports	2
Kinovea	X	✓	X	✓	X	✓	X	✓	Export as video, images, and CSV	110
LINCE PLUS	X	✓	✓	✓	✓	✓	✓	✓	Interactive charts; exports to data analysis software programs	3
Observer XT	✓	✓	✓	✓	✓	✓	X	✓	Export as reports, graphs, and charts, and CSV	38
SiliconCOACH	✓	✓	✓	✓	✓	✓	X	✓	Export as video and to Excel	4
Spark Motion Pro	✓	✓	X	✓	X	✓	✓	✓	Export as video, snapshot, and notes	1
Utilius Fairplay 5	✓	✓	✓	✓	X	?	X	✓	Export scenes	0

\* ✓ = yes; X = no; ? = not specified



### 3.3.1. How will educators annotate and analyze their videos?

Observers can annotate and analyze videos using isochronic, spatial and structured methods (Rolf *et al.*, 2014), which is supported by the 11 VATs. Key features include frame-by-frame analysis, split-screen playback, and comparison of student videos with expert demonstrations, facilitating visual feedback and imitation-based learning (White *et al.*, 2019). Additional features such as simultaneous footage, slow-motion, zoom, and frame-by-frame review enhance analysis. Observer XT (Post *et al.*, 2016) and SiliconCOACH (Carzoli *et al.*, 2022; Gabin *et al.*, 2012; Zimmerman *et al.*, 2009) offer these features, while LINC PLUS provides multi-video observation, device compatibility and adjustable playback speed (Soto *et al.*, 2019; Soto-Fernández *et al.*, 2021). Video annotation is time-consuming, as highlighted by White *et al.* (2019), who reported that coding one hour of surgical footage using Observer XT took three to five hours. This may be due to relying on mouse-based menu options. FEVA (Shrestha *et al.*, 2023) and LINC PLUS (Soto *et al.*, 2019; Soto-Fernández *et al.*, 2021) offer keyboard shortcuts, speed controls, and real-time labelling, improving efficiency. While most VATs support these features, it is unclear if Catapult does (Laughlin *et al.*, 2019). Live processing, as available in the tools such as Observer XT (Zimmerman *et al.*, 2009) and LINC PLUS (Soto *et al.*, 2019; Soto-Fernández *et al.*, 2021), could expedite annotation, although VATs such as Kinovea require post-processing (Carzoli *et al.*, 2022).

### 3.3.2. Will educators collaborate on their analyses?

Any VAT will suffice for single observer annotation. However, seven VATs support multiple observers, including Dartfish (Maykut *et al.*, 2015; Rucci & Tomporowski, 2010), LINC PLUS (Gabin *et al.*, 2012), Observer XT (Oliveira *et al.*, 2013) and SiliconCOACH (Shultz *et al.*, 2013), which allow reliability calculations between observers. LINC PLUS also supports collaborative work via mobile devices using QR codes (Soto *et al.*, 2019; Soto-Fernández *et al.*, 2021). Mobile VATs offer immediate feedback and quick video creation, with apps available for both Apple and Android, designed to operate seamlessly across various platforms. For example, Dartfish Express mirrors Dartfish, and Hudl Technique mirrors Hudl Sportscode (Laughlin *et al.*, 2019). Mobile apps of Catapult, CoachNow, LINC PLUS, and Spark Motion Pro also enhance versatility. Clear instructions on recording techniques, such as camera positioning and distance, are crucial for accurate observation and measurement (Ong *et al.*, 2015; Rucci & Tomporowski, 2010).

### 3.3.3. How much does it cost?

Detailed pricing information for most VATs is not publicly available. However, differences in access models can be identified. FEVA, Kinovea and LINC PLUS are free of charge, open-source VATs (LINC PLUS, 2025; Shrestha *et al.*, 2023). In contrast, CoachNow, Dartfish, SiliconCOACH and Spark Motion Pro require either a one-time purchase or a subscription plan, depending on additional features (CoachNow, n.d.; Dartfish, 2025; Siliconcoach, n.d.; SparkMotion, 2025). Pricing information for Catapult, Hudl Sportscode and Observer XT is not disclosed online (Catapult, n.d.; Hudl, 2007-2025; Noldus, 2025). While exact costs are difficult to determine, the accessibility and licensing model (free versus paid) can be considered a key factor in tool selection.

### 3.3.4. Who should upload the videos?

Responsibility for uploading videos varies, depending on the VAT and the educational context. In some platforms, such as CoachNow and Dartfish, students can upload their own recordings. In other cases, observers are responsible for recording and uploading performance footage, especially when institutional or privacy protocols limit student access. The ability to distribute software to students or export video analysis is advantageous. A digital drop box system allows fast and secure sharing of student footage with the observer (Hands *et al.*, 2009). Exporting data in a universal format such as .csv is valuable, as seen in LINC PLUS, CoachNow, Dartfish,

Kinovea, Observer XT and SiliconCOACH (CoachNow, n.d.; Dartfish, 2025; Gabin *et al.*, 2012; Kinovea, n.d.; LINC PLUS, 2025; Noldus, 2025; Siliconcoach, n.d.).

### **3.3.5. How secure are videos and reflections from unwanted viewing?**

Data management is crucial for storing student IDs, observer names, and other key details to analyze large groups. All VATs provide this functionality; it cannot be specified for Utilius Fairplay 5 (CCC Software, 2025). LINC PLUS also allows for storing essential characteristics, thus streamlining observational research (Gabin *et al.*, 2012).

### **3.3.6. Is the tool easy for educators to learn and use?**

Most VATs offer tutorials via YouTube or third-party creators, except FEVA, which has a demo and white paper on its website. Paid VATs provide comprehensive support through their websites, such as expert training sessions for CoachNow, Catapult and Spark Motion Pro. Observer XT includes customer support and tutorials via a MyNoldus account (Noldus, 2025), while Dartfish and Hudl Sportscode offer academy webinars (Dartfish, 2025; Hudl, n.d.). Kinovea provides multilingual tutorials and an active help forum, praised for ease of use and portability (Puig-Diví *et al.*, 2019). Utilius Fairplay 5 has a blog, a white paper, and a video analysis guide (CCC Software, 2025).

## **3.4. Review Findings**

The customizability of a VAT is a key consideration in nursing education, where tools must be adaptable to discipline-specific psychomotor skills. Among the 11 selected VATs, four stand out in this regard. LINC PLUS is open-source, highly adaptable, and supports fixed, mixed, and variable criteria coding (Gabin *et al.*, 2012; Prieto-Lage *et al.*, 2020; Soto-Fernández *et al.*, 2021). It accommodates direct observation, video analysis, and vocal feedback functionalities (Soto *et al.*, 2019; Soto-Fernández *et al.*, 2021). Kinovea, also open-source, offers features such as split-screen viewing and various annotation tools. It has been primarily used in research settings to measure parameters such as time, position, distance, angles, and both linear and angular kinematics (Kinovea, n.d.). While it supports a range of analytical functions, its customization options, such as the creation of new codes or commands, are currently more limited compared to some other tools. Dartfish provides advanced performance analysis options and supports the creation of media books that combine videos, images, annotations, and comments. It offers a high degree of customizability, although it requires a commercial license (Dartfish, 2025; Hands *et al.*, 2009). Observer XT is a comprehensive tool for behavioral research, allowing detailed coding and statistical analysis. It also supports integration with media and external data modules, although these modules may require additional clarification regarding price and setup (Dove & Astell, 2019; Noldus, 2025).

## **4. DISCUSSION and CONCLUSION**

This scoping review examines VATs for annotating psychomotor performances (RQ1) and customizing PNS annotations in nursing education (RQ2). The literature search identified 528 records, leading to 86 reports describing 93 tools. Ultimately, 48 peer-reviewed articles describe 18 VATs eligible for data extraction. Factors considered include language, open-source availability, footage, and annotation type. Mandatory features, such as multi-video angles for simultaneous analysis and isochronic, spatial, and structured annotation, led to the exclusion of seven additional VATs during the second screening. In the final phase, the strengths and weaknesses of 11 selected VATs were described, with four tools standing out for their customizability to nursing education.

To contextualize the findings, some observations should be mentioned. First, while 93 tools were initially eligible for extraction, 45 were no longer available, highlighting the rapidly evolving nature of technology in this field. Second, limited information was obtained from selected articles, as most authors only mentioned VATs in the method section without

discussing their advantages or disadvantages. Primary sources of information were VATs' websites and review articles. While the initial search strings addressed RQ1, an additional PubMed search using VAT names yielded 10 relevant studies discussing the strengths and weaknesses to address RQ2. Last, despite conducting the scoping review across 11 databases and a Google search, the information obtained was insufficient for a comprehensive understanding of VATs' applicability for providing feedback on PNS. This was particularly evident for non-open-source VATs, which required payment to download and test. For VATs such as Catapult, discovered via Google search, only limited website-provided information was available.

These findings highlight a mismatch between the availability of VATs and their applicability to the nursing education context. While many tools exist, few meet the criteria needed to support feedback on PNS performance. This underscores the need for tool selection to be guided not merely by availability or popularity, but by pedagogical fit, particularly the ability to deliver isochronic, spatial, and structured feedback. For lecturers, selecting a VAT that aligns with these feedback types supports student learning effectively. The review also reveals a significant gap in the literature, the lack of peer-reviewed evaluations of VATs in nursing education. Most studies did not address usability, educational impact, or implementation feasibility, which points to the need for further empirical research. In addition, the frequent discontinuation of tools suggests that institutions may benefit from investing in open-source or customizable platforms to ensure long-term access and adaptability.

This review provides insights into VATs in international literature and identifies those that can be customized for feedback to nursing students on their PNS performance. Based on our scoping review, the open-source VATs LINC PLUS and Kinovea, and paid VATs Dartfish and Observer XT meet all required criteria, offering multi-video annotation, various annotation types and adaptability for PNS assessment. They stand out for their flexibility, versatility, and collaborative and data management options. This scoping review not only maps the current VAT landscape but also serves as a springboard for the development of a tailored VAT for feedback on PNS. By identifying critical functional requirements and existing gaps, this review can inform the design of a next-generation VAT that integrates the strengths of existing tools while addressing the specific pedagogical and practical needs of nursing education. Such a tool could enhance student learning outcomes and offer a time-efficient, user-friendly solution for educators in higher education.

As an initial mapping of this field, this review also highlights the need for further empirical research to evaluate how specific VAT features impact feedback quality, learning effectiveness, and implementation feasibility in nursing curricula.

### Acknowledgments

The authors wish to thank Thomas Vandendriessche, Anouk D'Hont, Norin Hamouda, Krizia Tuand and Chayenne Van Meel, the reference librarians of KU Leuven Libraries – 2Bergen (Leuven, Belgium), for their help in conducting the systematic literature search.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

### Contribution of Authors

**Greet Leysens** and **Rani Claus** are the corresponding authors of the article. **Greet Leysens:** Design of methodology, Conceptualization, Investigation, Resources, Writing - original draft, Writing - review and editing. **Rani Claus:** Conceptualization, Investigation, Resources, Writing – original draft, Writing – review. **Wim Van Petegem** and **Nathalie Charlier:** Supervision, Writing – review.

## Orcid

Greet Leysens  <https://orcid.org/0000-0002-3132-3238>

Rani Claus  <https://orcid.org/0000-0001-6700-9665>

Wim Van Petegem  <https://orcid.org/0000-0002-4553-4407>

Nathalie Charlier  <https://orcid.org/0000-0002-9511-956X>

## REFERENCES

- Abdelrasoul, E., Mahmoud, I., Stergiou, P., & Katz, L. (2015). The accuracy of a real time sensor in an instrumented basketball. *Procedia Engineering*, 112, 202-206. <https://doi.org/10.1016/j.proeng.2015.07.200>
- Amri-Dardari, A., Mkaouer, B., Amara, S., Hammoudi-Nassib, S., Habacha, H., & Zohra BenSalah, F. (2022). Immediate effect of self-Modelling with internal versus external focus of attention on teaching/learning gymnastics motor-skills. *Journal of Human Kinetics*, 84(1), 224-232. <https://doi.org/10.2478/hukin-2022-0103>
- Amri-Dardari, A., Mkaouer, B., Nassib, S.H., Amara, S., Amri, R., & Ben Salah, F.Z. (2020). The effects of video modeling and simulation on teaching / learning basic vaulting jump on the vault table. *Science of Gymnastics Journal*, 12(3), 325-344. <https://doi.org/10.52165/sgj.12.3.325-344>
- Andrews, B.S., & Bressan, E.S. (2018). The effect of synchronised metronome training: a case study in a single leg, below knee Paralympic sprinter. *African journal of Disability*, 7(1), 1-6. <https://doi.org/10.4102/ajod.v7i0.367>
- Ardley, J., & Hallare, M. (2020). The feedback cycle: Lessons learned with video annotation software during student teaching. *Journal of Educational Technology Systems*, 49(1), 94-112. <https://doi.org/10.1177/0047239520912343>
- Ardley, J., & Johnson, J. (2019). Video annotation software in teacher education: Researching university supervisor's perspective of a 21st-century technology. *Journal of Educational Technology Systems*, 47(4), 479-499. <https://doi.org/10.1177/0047239518812715>
- Aubert, O., Prié, Y., & Canellas, C. (2014). Leveraging video annotations in video-based e-learning. *arXiv preprint arXiv:1404.4607*. <https://doi.org/10.48550/arxiv.1404.4607>
- Bahula, T., & Kay, R. (2021). Exploring student perceptions of video-based feedback in higher education: A systematic review of the literature. *Journal of Higher Education Theory and Practice*, 21(4), 248-258. <https://doi.org/10.33423/jhetp.v21i4.4224>
- Barris, S., & Button, C. (2008). A review of vision-based motion analysis in sport. *Sports Medicine*, 38(12), 1025-1043. <https://doi.org/10.2165/00007256-200838120-00006>
- Beseler, B., Plumb, M.S., Spittle, M., Johnson, N.F., Harvey, J.T., & Mesagno, C. (2024). Examining single session peer-teaching instructional approaches on pre-service physical education teachers' throwing techniques. *Perceptual and Motor Skills*, 131(1), 246-266. <https://doi.org/10.1177/00315125231214126>
- Bobo, L., Benson, A.A., & Green, M. (2012). The effect of self-reported efficacy on clinical skill performance. *Athletic Training Education Journal*, 7(4), 176-186. <https://scholarworks.sfasu.edu/kinesiology/27>
- Cabarkapa, D., Fry, A.C., Cabarkapa, D.V., Myers, C.A., Jones, G.T., & Deane, M.A. (2021). Kinetic and kinematic characteristics of proficient and non-proficient 2-point and 3-point basketball shooters. *Sports (Basel)*, 10(1), 2. <https://doi.org/10.3390/sports10010002>
- Carzoli, J.P., Sousa, C.A., Helms, E.R., & Zourdos, M.C. (2022). Agreement between Kinovea video analysis and the open barbell system for resistance training movement outcomes. *Journal of human kinetics*, 81(1), 27-39. <https://doi.org/10.2478/hukin-2022-0003>
- Cassano, G., & Di Blas, N. (2023). A tool to support students-to-teacher feedback in asynchronous online contexts. *IEEE Transactions on Learning Technologies*, 17, 585-593. <https://doi.org/10.1109/TLT.2023.3273109>



- Catapult. (n.d.). *Sports video analysis software*. <https://www.catapult.com/solutions/video-analysis>
- CCC Software. (2025). *Utilius fairplay 5*. <https://ccc-sportsoftware.de/en/produkte/utillius-fairplay-5/>
- Chiappedi, M., Togni, R., De Bernardi, E., Baschenis, I.M.C., Battezzato, S., Balottin, U., Toffola, E.D., & Bejor, M. (2012). Arm trajectories and writing strategy in healthy children. *BMC Pediatrics*, 12(1), 173-173. <https://doi.org/10.1186/1471-2431-12-173>
- Chorney, J.M.L., & Kain, Z.N. (2009). Behavioral analysis of children's response to induction of anesthesia. *Anesthesia and analgesia*, 109(5), 1434-1440. <https://doi.org/10.1213/ane.0b013e3181b412cf>
- CoachNow. (n.d.). *Welcome to ConnectedCoaching*. <https://coachnow.io/>
- Dadashi, F., Crettenand, F., Millet, G.P., Seifert, L., Komar, J., & Aminian, K. (2013). Automatic front-crawl temporal phase detection using adaptive filtering of inertial signals. *Journal of Sports Sciences*, 31(11), 1251-1260. <https://doi.org/10.1080/02640414.2013.778420>
- Dartfish. (2025). *Video analysis solutions for sports performance*. <https://www.dartfish.com/>
- Douglas, K.A., Lang, J., & Colasante, M. (2014). The challenges of blended learning using a media annotation tool. *Journal of University Teaching & Learning Practice*, 11(2). <https://doi.org/10.53761/1.11.2.7>
- Dove, E., & Astell, A.J. (2019). Kinect Project: People with dementia or mild cognitive impairment learning to play group motion-based games. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5(1), 475-482. <https://doi.org/10.1016/j.trci.2019.07.008>
- Earp, J.E., Newton, R.U., Cormie, P., & Blazevich, A.J. (2016). Faster movement speed results in greater tendon strain during the loaded squat exercise. *Frontiers in Physiology*, 7, 366-366. <https://doi.org/10.3389/fphys.2016.00366>
- Epstein, I., Baljko, M., Thumlert, K., Kelly, E., Smith, J.A., Su, Y., Zaki-Azat, J., & May, N.M. (2020). "A video of myself helps me learn": A scoping review of the evidence of video-making for situated learning. *International Journal for the Scholarship of Teaching and Learning*, 14(1), 9. <https://doi.org/10.20429/ijstol.2020.140109>
- Evi-Colombo, A., Cattaneo, A., & Bétrancourt, M. (2020). Technical and pedagogical affordances of video annotation: A literature review. *Journal of Educational Multimedia and Hypermedia*, 29(3), 193-226. <https://www.learntechlib.org/primary/p/215718/>.
- Frehner, E., Tulloch, A., & Glaister, K. (2012). "Mirror, mirror on the wall": The power of video feedback to enable students to prepare for clinical practice. In Herrington, A., Schrape, J., & Singh, K. (Eds.), *Engaging students with learning technologies* (1-14). Curtin University.
- Gabin, B., Camerino, O., Anguera, M.T., & Castañer, M. (2012). Lince: Multiplatform sport analysis software. *Procedia - Social and Behavioral Sciences*, 46, 4692-4694. <https://doi.org/10.1016/j.sbspro.2012.06.320>
- Gonzalvo, A.R.A., Esparza Yáñez, D., Tricás Moreno, J.M., & Lucha López, M.O. (2017). Validation of a force platform clinical for the assessment of vertical jump height. *Journal of Human Sport and Exercise*, 12(2), 367-379. <https://doi.org/10.14198/jhse.2017.122.13>
- Hands, B.P., Parker, H., Coffey, A., Clark-Burg, K., Das, A., Gerrard, P., Hackett, C., Jenkins, S., MacNish, J., & Miller, D. (2009). *The PHENC Project final report: Interactive video analysis to develop learning and assessment of university students' practical and communication skills*. [https://ltr.edu.au/resources/CG7-385\\_Notre%20Dame\\_Hands\\_Final%20Report\\_v2\\_May09.pdf](https://ltr.edu.au/resources/CG7-385_Notre%20Dame_Hands_Final%20Report_v2_May09.pdf)
- Hudl. (n.d.). *Hudl Sportscode fully customizable performance analysis*. [https://www.hudl.com/en\\_gb/products/sportscode](https://www.hudl.com/en_gb/products/sportscode)
- INACSL. (2011). Standard I: Terminology. *Clinical Simulation in Nursing*, 7(4), S3-S7. <https://doi.org/10.1016/j.ecns.2011.05.005>



- Ishac, K., & Eager, D. (2021). Evaluating martial arts punching kinematics using a vision and inertial sensing system. *Sensors*, 21(6), 1-25. <https://doi.org/10.3390/s21061948>
- Judge, L.W., Hunter, I., & Gilreath, E. (2008). Using sport science to improve coaching: A case study of the American record holder in the women's hammer throw. *International Journal of Sports Science & Coaching*, 3(4), 477-488. <https://doi.org/10.1260/174795408787186440>
- Kinovea. (n.d.). *A microscope for your videos* <https://www.kinovea.org/>
- Lago-Fuentes, C., Rey, E., Padrón-Cabo, A., Sal de Rellán-Guerra, A., Fragueiro-Rodríguez, A., & García-Núñez, J. (2018). Effects of core strength training using stable and unstable surfaces on physical fitness and functional performance in professional female futsal players. *Journal of human kinetics*, 65(1), 213-224. <https://doi.org/10.2478/hukin-2018-0029>
- Lam, C., & Habil, H. (2021a). Enriching student learning through video-annotated peer feedback activity: A guide. *International Journal of Academic Research in Progressive Education and Development*, 10(3), 46-60. <http://dx.doi.org/10.6007/IJARPED/v10-i3/10712>
- Lam, C., & Habil, H. (2021b). The use of video annotation in education: A review. *Asian Journal of University Education*, 17(4), 84-94. <https://doi.org/10.24191/ajue.v17i4.16208>
- Laughlin, M.K., Hodges, M., & Irraggi, T. (2019). Deploying video analysis to boost instruction and assessment in physical education. *Journal of Physical Education, Recreation & Dance*, 90(5), 23-29. <https://doi.org/10.1080/07303084.2019.1580637>
- Leysens, G., Claus, R., Van Petegem, W., & Charlier, N. (December 20, 2024). *Video annotation tools to improve psychomotor skills in health professional education: a scoping review protocol* (Version V1) KU Leuven RDR. <https://doi.org/10.48804/PWWPOD>
- Liebermann, D.G., Katz, L., Hughes, M.D., Bartlett, R.M., McClements, J., & Franks, I.M. (2002). Advances in the application of information technology to sport performance. *Journal of Sports Sciences*, 20(10), 755-769. <https://doi.org/10.1080/026404102320675611>
- LINCE PLUS. (2025). *LINCE PLUS sports training, reimagined*. <https://lince-plus.com/>
- Loukas, C., Gazis, A., & Kanakis, M.A. (2020). Surgical performance analysis and classification based on video annotation of laparoscopic tasks. *Journal of the Society of Laparoendoscopic Surgeons*, 24(4), e2020.00057. <https://doi.org/10.4293/JSLs.2020.00057>
- Martins, J., Baptista, R., Coutinho, V., Fernandes, M., & Fernandes, A. (2018). *Simulation in nursing and midwifery education*. WHO. <https://iris.who.int/bitstream/handle/10665/345156/WHO-EURO-2018-3296-43055-60253-eng.pdf?sequence=2&isAllowed=y>
- Mayer, R.E. (2014). *The Cambridge handbook of multimedia learning*, edited by Richard E. Mayer, University of California, Santa Barbara (Second Edition ed.). Cambridge University Press.
- Maykut, J.N., Taylor-Haas, J.A., Paterno, M.V., DiCesare, C.A., & Ford, K.R. (2015). Concurrent validity and reliability of 2d kinematic analysis of frontal plane motion during running. *International journal of Sports Physical Therapy*, 10(2), 136-146.
- McDonald, D.A., Delgadillo, J.Q., Fredericson, M., McConnell, J., Hodgins, M., & Besier, T.F. (2011). Reliability and accuracy of a video analysis protocol to assess core ability. *PM & R: the Journal of Injury, Function, and Rehabilitation*, 3(3), 204-211. <https://doi.org/10.1016/j.pmrj.2010.12.007>
- McFadden, J., Ellis, J., Anwar, T., & Roehrig, G. (2014). Beginning science teachers' use of a digital video annotation tool to promote reflective practices. *Journal of Science Education and Technology*, 23(3), 458-470. <https://doi.org/10.1007/s10956-013-9476-2>
- Miles, D.A. (2018). Simulation learning and transfer in undergraduate nursing education: A grounded theory study. *Journal of Nursing Education*, 57(6), 347-353. <https://doi.org/10.3928/01484834-20180522-05>
- Myer, G.D., Ford, K.R., Brent, J.L., & Hewett, T.E. (2012). An integrated approach to change the outcome part I: Neuromuscular screening methods to identify high ACL injury risk athletes. *Journal of strength and conditioning research*, 26(8), 2265-2271. <https://doi.org/10.1519/JSC.0b013e31825c2b8f>

- Nagel, I., & Engeness, I. (2021). Peer feedback with video annotation to promote student teachers' reflections. *Acta Didactica Norden*, 15(3), 24. <https://doi.org/10.5617/ADNO.8192>
- Noldus. (2025). *The Observer XT*. <https://www.noldus.com/observer-xt>
- O'Donoghue, P., & Holmes, L. (2014). *Data analysis in sport*. Routledge.
- Oliveira, A., Pinho, C., Monteiro, S., Marcos, A., & Marques, A. (2013). Usability testing of a respiratory interface using computer screen and facial expressions videos. *Computers in Biology and Medicine*, 43(12), 2205-2213. <https://doi.org/10.1016/j.combiomed.2013.10.010>
- Ong, N.T., Lohse, K.R., & Hodges, N.J. (2015). Manipulating target size influences perceptions of success when learning a dart-throwing skill but does not impact retention. *Frontiers in Psychology*, 6, 1378-1378. <https://doi.org/10.3389/fpsyg.2015.01378>
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, 5, 1-10. <https://doi.org/10.1186/s13643-016-0384-4>
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. <https://doi.org/10.1136/bmj.n71>
- Peeters, A., Carling, C., Piscione, J., & Lacome, M. (2019). In-match physical performance fluctuations in international rugby sevens competition. *Journal of Sports Science & Medicine*, 18(3), 419-426.
- Pérez-Torregrosa, A.B., Díaz-Martín, C., & Ibáñez-Cubillas, P. (2017). The use of video annotation tools in teacher training. *Procedia - Social and Behavioral Sciences*, 237, 458-464. <https://doi.org/10.1016/j.sbspro.2017.02.090>
- Peters, M., Godfrey, C., McInerney, P., Munn, Z., Tricco, A., & Khalil, H. (2024). JBI Manual for evidence synthesis chapter 10 Scoping review. *JBI*. <https://doi.org/10.46658/JBIMES-24-09>
- Polanin, J.R., Pigott, T.D., Espelage, D.L., & Grotzinger, J.K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, 10(3), 330-342. <https://doi.org/10.1002/jrsm.1354>
- Post, P.G., Aiken, C.A., Laughlin, D.D., & Fairbrother, J.T. (2016). Self-control over combined video feedback and modeling facilitates motor learning. *Human Movement Science*, 47, 49-59. <https://doi.org/10.1016/j.humov.2016.01.014>
- Prieto-Lage, I., Rodríguez-Souto, M., Prieto, M.A., & Gutiérrez-Santiago, A. (2020). Technical analysis in Tsuri-goshi through three complementary observational analysis. *Physiology & Behavior*, 216, 112804. <https://doi.org/10.1016/j.physbeh.2020.112804>
- Puig-Diví, A., Escalona-Marfil, C., Padullés-Riu, J.M., Busquets, A., Padullés-Chando, X., & Marcos-Ruiz, D. (2019). Validity and reliability of the Kinovea program in obtaining angles and distances using coordinates in 4 perspectives. *PLoS One*, 14(6), e0216448. <https://doi.org/10.1371/journal.pone.0216448>
- Puklavec, A., Antekolovic, L., & Mikulic, P. (2021). Acquisition of the long jump skill using varying feedback. *Croatian Journal of Education-Hrvatski Casopis Za Odgoj I Obrazovan je*, 23(1), 107-132. <https://doi.org/10.15516/cje.v23i1.3994>
- Raiola, G., Parisi, F., Giugno, Y., & Di Tore, P.A. (2013). Video analysis applied to volleyball didactics to improve sport skills. *Journal of Human Sport and Exercise*, 8(Proc2), 307-313. <https://doi.org/10.4100/jhse.2012.8.Proc2.33>
- Rich, P.J., & Trip, T. (2011). Ten essential questions educators should ask when using video annotation tools. *TechTrends*, 55, 16-24. <https://doi.org/10.1007/s11528-011-0537-1>

- Rolf, R., Reuter, H., Abel, M., & Kai-Christoph, H. (2014). Requirements of students for video-annotations in lecture recordings. *Interactive Technology and Smart Education*, 11(3), 223-234. <https://doi.org/10.1108/ITSE-07-2014-0021>
- Rucci, J.A., & Tomporowski, P.D. (2010). Three types of kinematic feedback and the execution of the hang power clean. *Journal of Strength and Conditioning Research*, 24(3), 771-778. <https://doi.org/10.1519/JSC.0b013e3181cbab96>
- Sain, D. (2022). *The Use of Video Analysis to Improve Performance* [Doctor of Education Dissertations 108, Gardner-Webb University]. Digital Commons@Gardner-Webb University <https://digitalcommons.gardner-webb.edu/education-dissertations/108>
- Schärer, C., Huber, S., Bucher, P., Capelli, C., & Hübner, K. (2021). Maximum strength benchmarks for difficult static elements on rings in male elite gymnastics. *Sports (Basel)*, 9(6), 78. <https://doi.org/10.3390/sports9060078>
- Schulz, D., & Gaudreault, K. (2023). GoReact: Video annotation software to foster feedback in physical education instruction. *Strategies*, 36(3), 8-13. <https://doi.org/10.1080/08924562.2023.2195453>
- Shih, H.C. (2018). A survey of content-Aware video analysis for sports. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(5), 1212-1231. <https://doi.org/10.1109/tcsvt.2017.2655624>
- Shorey, S., Chan, V., Rajendran, P., & Ang, E. (2021). Learning styles, preferences and needs of generation Z healthcare students: Scoping review. *Nurse Education in Practice*, 57, 103247. <https://doi.org/10.1016/j.nepr.2021.103247>
- Shrestha, S., Sentosatio, W., Peng, H., Fermuller, C., & Aloimonos, Y. (2023). FEVA: Fast Event Video Annotation tool. *arXiv preprint arXiv:2301.00482*. <https://doi.org/10.48550/arxiv.2301.00482>
- Shultz, R., Anderson, S.C., Matheson, G.O., Marcello, B., & Besier, T. (2013). Test-retest and interrater reliability of the functional movement screen. *Journal of Athletic Training*, 48(3), 331-336. <https://doi.org/10.4085/1062-6050-48.2.11>
- Siliconcoach. (n.d.). *Siliconcoach Pro: Analyse performance in detail*. <https://www.siliconcoach.com/siliconcoachPro>
- Soto, A., Camerino, O., Iglesias, X., Anguera, M.T., & Castañer, M. (2019). LINCE PLUS: Research software for behavior video analysis. *Apunts. Educación Física y Deportes*, 3(137), 149-153. [https://doi.org/10.5672/apunts.2014-0983.es.\(2019/3\).137.11](https://doi.org/10.5672/apunts.2014-0983.es.(2019/3).137.11)
- Soto-Fernández, A., Camerino, O., Iglesias, X., Anguera, M.T., & Castañer, M. (2021). LINCE PLUS software for systematic observational studies in sports and health. *Behavior Research Methods*, 1-9. <https://doi.org/10.3758/s13428-021-01642-1>
- Souissi, H., Souissi, M.A., Trabelsi, O., Ben Chikha, A., Gharbi, A., & Souissi, N. (2023). Peer-to-peer online video feedback with pedagogical activity improves the snatch learning during the COVID-19-induced confinement in young weightlifting athletes. *International journal of Sports Science & Coaching*, 18(6), 2151-2159. <https://doi.org/10.1177/17479541221122385>
- Souissi, M.A., Ammar, A., Trabelsi, O., Glenn, J.M., Boukhris, O., Trabelsi, K., Bouaziz, B., Zmijewski, P., Souissi, H., Chikha, A.B., Driss, T., Chtourou, H., Hoekelmann, A., & Souissi, N. (2021). Distance motor learning during the covid-19 induced confinement: Video feedback with a pedagogical activity improves the snatch technique in young athletes. *International journal of Environmental Research and Public Health*, 18(6), 1-13. <https://doi.org/10.3390/ijerph18063069>
- SparkMotion. (2025, January 8). *Motion Analysis for your business has never been so easy*. <https://sparkmotion.com/>
- Ste-Marie, D.M., Carter, M.J., Law, B., Vertes, K., & Smith, V. (2016). Self-controlled learning benefits: exploring contributions of self-efficacy and intrinsic motivation via path analysis. *Journal of Sports Sciences*, 34(17), 1650-1656. <https://doi.org/10.1080/02640414.2015.1130236>

- Ste-Marie, D.M., Vertes, K.A., Law, B., & Rymal, A.M. (2012). Learner-controlled self-observation is advantageous for motor skill acquisition. *Frontiers in Psychology*, 3, 556. <https://doi.org/10.3389/fpsyg.2012.00556>
- Tay, D.L., Ellington, L., Towsley, G.L., Supiano, K., & Berg, C.A. (2021). Emotional expression in conversations about advance care planning among older adult home health patients and their caregivers. *Patient Education and Counseling*, 104(9), 2232-2239. <https://doi.org/10.1016/j.pec.2021.02.029>
- Tayech, A., Mejri, M.A., Makhoul, I., Uthoff, A., Hambli, M., Behm, D.G., & Chaouachi, A. (2022). Reliability, criterion-concurrent validity, and construct-discriminant validity of a head-marking version of the taekwondo anaerobic intermittent kick test. *Biology of Sport*, 39(4), 951-963. <https://doi.org/10.5114/biolsport.2022.109459>
- Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D.J., Horsley, T., Weeks, L., Hempel, S., Akl, E.A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M.G., Garritty, C., ... Straus, S.E. (2018). PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467-473. <https://doi.org/10.7326/m18-0850>
- Vermeulen, M. (2019). *Stappenplannen. Technische verpleegkundige verstrekingen* (12th ed.). Bohn Stafleu Van Loghum.
- Walker, S.G., Mattson, S.L., & Sellers, T.P. (2020). Increasing accuracy of rock-climbing techniques in novice athletes using expert modeling and video feedback. *Journal of Applied Behavior Analysis*, 53(4), 2260-2270. <https://doi.org/10.1002/jaba.694>
- White, E.J., McMahon, M., Walsh, M.T., Coffey, J.C., & O'Sullivan, L.W. (2019). A study of laparoscopic instrument use during colorectal surgery. *Applied Ergonomics*, 78, 301-308. <https://doi.org/10.1016/j.apergo.2018.02.010>
- Yang, K., Jin, X., Wang, Z., Fang, Y., Li, Z., Yang, Z., Cong, J., Yang, Y., Huang, Y., & Wang, L. (2022). Robot-assisted subretinal injection system: development and preliminary verification. *BMC Ophthalmology*, 22(1), 484-484. <https://doi.org/10.1186/s12886-022-02720-4>
- Zimmerman, P.H., Bolhuis, J.E., Willemsen, A., Meyer, E.S., & Noldus, L.P.J.J. (2009). The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods*, 41(3), 731-735. <https://doi.org/10.3758/brm.41.3.731>

## APPENDIX

### Appendix I: Inclusion and exclusion criteria.

	Description	Inclusion criteria	Exclusion criteria
Participants	Participants in included studies receive feedback on their execution of psychomotor performance.	<ul style="list-style-type: none"> <li>• focus on psychomotor learning domain</li> <li>• persons executing a movement-oriented performance</li> </ul>	<ul style="list-style-type: none"> <li>• focus on cognitive and affective learning domain, such as feedback on communication skills or knowledge</li> <li>• persons executing other than movement-oriented performances</li> </ul>
Concept	<p>Studies that explore features of VATs capable of generating isochronic, spatial, and structured annotations on videos of psychomotor performances.</p> <p>Information on availability, accessibility, financial feasibility, and customizability of these tools will also be collected.</p>	<p>VAT</p> <ul style="list-style-type: none"> <li>• makes annotations on videos of movement-orientated performances</li> <li>• is available</li> <li>• is accessible</li> <li>• is customizable</li> </ul>	<p>VAT</p> <ul style="list-style-type: none"> <li>• has focus on cognitive and affective learning domain</li> <li>• for machine learning algorithm or automated image annotation</li> <li>• for real-time video analytics, such as wearable sensors, or trackers</li> <li>• for measurements on medical images such as MRI</li> <li>• is no longer available</li> <li>• is not customizable</li> </ul>
Context	<p>Studies that apply VATs in contexts transferable to health professional education will be included.</p> <p>Given that PNS involve physical, movement-oriented performances, and considering the extensive use of VAT in sports sciences, a broader scope on VATs applications will be adopted for inspiration.</p> <p>Focusing initially on the psychomotor aspects of a new PNS is a powerful learning strategy, fostering students' progression from novice to expert. Video footage, paired with concrete feedback on a context-free PNS, will enhance and</p>	<p>Use of a VAT in health professional education to provide feedback to enhance performance of PNS in students in their novice learning phase.</p>	<ul style="list-style-type: none"> <li>• feedback on performances related to cognitive or attitude learning domain</li> <li>• feedback related to simulation-based skills or critical reasoning</li> </ul>



	accelerate students' mastery of task performances.		
Type of sources	<p>Primary research and additionally, systematic reviews, meta-analyses, text and opinion papers, guidelines, and conference proceedings will be considered for inclusion.</p> <p>Articles must be available in full text and published in Dutch, English, French, or German. No restrictions on the publication date will be applied.</p>	<ul style="list-style-type: none"> <li>• information in Dutch, English, French and German</li> <li>• primary research studies</li> <li>• systematic reviews</li> <li>• meta-analyses</li> <li>• text and opinion papers</li> <li>• guidelines</li> <li>• conference proceedings</li> </ul>	<ul style="list-style-type: none"> <li>• information in language other than Dutch, English, French and German</li> <li>• no full text available</li> <li>• abstract conferences</li> </ul>

**Appendix II: Search Strings and hits.**

Database	String	Hits (dated 5/2/2024)
PubMed	#1: (("Videotape Recording"[Mesh] OR "video annotat*"[tiab] OR "annotated video"[tiab:~2] OR "annotation video"[tiab:~2] OR "annotated videos"[tiab:~2] OR "annotation videos"[tiab:~2] OR "annotations video"[tiab:~2] OR "annotations videos"[tiab:~2] OR "video analys*"[tiab]) AND ("Knowledge of Results, Psychological"[Mesh] OR "Formative Feedback"[Mesh] OR "Peer Review"[Mesh] OR ("Learning"[Mesh:NoExp] AND "2010:2015"[mhda]) OR feedback[tiab])) AND ("Psychomotor Performance"[Mesh] OR "Athletic Performance"[Mesh] OR "Physical Functional Performance"[Mesh] OR "Psychomotor Performance"[tiab] OR "Athletic Performance"[tiab] OR "motor performance"[tiab] OR "psychomotor coordination"[tiab] OR "motor coordination"[tiab] OR "physical performance"[tiab] OR "motor skill*"[tiab] OR "task performance"[tiab])	46
	#2: (((("Videotape Recording"[Mesh] OR "video annotat*"[tiab] OR "annotated video"[tiab:~2] OR "annotation video"[tiab:~2] OR "annotated videos"[tiab:~2] OR "annotation videos"[tiab:~2] OR "annotations video"[tiab:~2] OR "annotations videos"[tiab:~2] OR "video analys*"[tiab]) AND ("Knowledge of Results, Psychological"[Mesh] OR "Formative Feedback"[Mesh] OR "Peer Review"[Mesh] OR ("Learning"[Mesh:NoExp] AND "2010:2015"[mhda]) OR "feedback"[tiab])) OR ("video feedback"[tiab])) AND ("Psychomotor Performance"[Mesh] OR "Athletic Performance"[Mesh] OR "Physical Functional Performance"[Mesh] OR "Psychomotor Performance"[tiab] OR "Athletic Performance"[tiab] OR "motor performance"[tiab] OR "psychomotor coordination"[tiab] OR "motor coordination"[tiab] OR "physical performance"[tiab] OR "motor skill*"[tiab] OR "task performance"[tiab])	87
	#3 ( ("Videotape Recording"[Mesh] OR "video annotat*"[tiab] OR "annotated video"[tiab:~2] OR "annotation video"[tiab:~2] OR "annotated videos"[tiab:~2] OR "annotation videos"[tiab:~2] OR "annotations video"[tiab:~2] OR "annotations videos"[tiab:~2] OR "video analys*"[tiab]) AND ("Knowledge of Results, Psychological"[Mesh] OR "Formative Feedback"[Mesh] OR "Formative Feedback" [tiab] OR "Peer Review"[Mesh] OR "Peer Review"[tiab] OR ("Learning"[Mesh:NoExp] AND "2010:2015"[mhda]) OR "feedback"[tiab] OR "Self Efficacy"[Mesh] OR "Self Efficacy"[tiab])) AND ("Psychomotor Performance"[Mesh] OR "Athletic Performance"[Mesh] OR "Physical Functional Performance"[Mesh] OR "Psychomotor Performance"[tiab] OR "Athletic Performance"[tiab] OR "motor performance"[tiab] OR "psychomotor coordination"[tiab] OR "motor coordination"[tiab] OR "physical performance"[tiab] OR "motor skill*"[tiab] OR "task performance"[tiab])	51
Web of Science	#1 (("video annotation" OR "video annotation software" OR "video annotation*" OR "video analysis*") AND ("feedback" OR "learning" OR "training" OR "professional development" OR "video assisted learning" OR "coaching") AND ("Psychomotor Performance" OR "psychomotor skill" OR "motor skill" OR "nursing skill" OR "sport performance"))	22
	#2 ((ALL=("performance")) OR ALL=("motor skill")) AND ALL=("video annotat*")	224

CINAHL	#1 TX "skill performance" AND TX "video feedback"	5
	#2 TX "video analysis" AND TX feedback AND TX performance	62
Scopus	#1 TITLE-ABS-KEY ("skill performance") AND TITLE-ABS-KEY ("video feedback")	8
	#2 TITLE-ABS-KEY ("video annotation") AND TITLE-ABS-KEY ("performance") AND TITLE-ABS-KEY ("feedback")	20
EuropePMC	#1 "skill performance" AND "video feedback"	43
	#2 "video analysis software" AND feedback AND performance	102
CENTRAL via Cochrane Library	#1 "skill performance" AND "video feedback"	1
	#2 "video annotation" AND "performance" AND "feedback"	1
ERIC	#1 "Video Technology" AND "psychomotor skill" AND "Feedback (Response)"	16
	#2 "Video Technology" AND "psychomotor skill" AND Coaching (Performance)	4
FiS Bildung	#1 ( ( ( ( ( ( ( (Subject: "VIDEO ANALYSIS SOFTWARE") or (Subject: "VIDEO ANNOTATION TOOLS") ) or (Subject: "VIDEO ANALYSIS") ) or (Subject: "VIDEO ANNOTATION") ) and (Subject: "SKILL ACQUISITION") ) or (Subject: "SKILL TRAINING") ) or (Subject: "SKILL DEVELOPMENT") ) and (Subject: FEEDBACK) ) or (Subject: "FEED FORWARD")	229
SportDiscus	#1 TX "skill performance" AND TX "video feedback"	4
	#2 TX "video analysis" AND TX feedback AND TX performance	24
IEEE	#1 ("Full Text & Metadata": "skill performance") AND ("Full Text & Metadata": "video feedback")	6
	#2 ("Full Text & Metadata": "video analysis") AND ("Full Text & Metadata": "feedback") AND ("Full Text & Metadata": "skill performance")	4
ACM digital library	#1 [All: "skill performance"] AND [All: "video feedback"]	2
	#2 [All: "video analysis"] AND [All: "feedback"] AND [All: "motor skill"]	38

## Deficit Thinking Scale for teachers: A validity and reliability study in Turkish context

Abide Ocak<sup>1</sup>, İsmail Çimen<sup>2\*</sup>

<sup>1</sup>Ministry of National Education, Bursa, Türkiye

<sup>2</sup>Bursa Uludağ University, Faculty of Education, Department of Educational Sciences, Bursa, Türkiye

### ARTICLE HISTORY

Received: Jan. 3, 2025

Accepted: Sep. 19, 2025

### Keywords:

Deficit thinking,  
Deficit theory,  
Scale development.

**Abstract:** Deficit thinking is one of the theories developed to explain achievement gaps among different student groups. It attributes academic failure to perceived deficiencies within students, disregarding the role of structural inequalities in education. This study aims to develop a valid and reliable scale to measure Turkish teachers' attitudes toward deficit thinking, which has not previously been operationalized in Türkiye. Based on a rigorous scale development process, including literature review, expert consultation, pilot testing, exploratory and confirmatory factor analyses, the final version of the scale consists of 22 items and five dimensions. Exploratory factor analysis ( $n = 323$ ; KMO = .87; Bartlett's  $\chi^2 = 4983.99$ ,  $p < .001$ ) revealed a five-factor structure (Blaming the Environment, Educability, Oppression, Blaming the Victim, and Pseudoscience) that explained 66.13% of the total variance. Confirmatory factor analysis ( $n = 569$ ) supported this structure with good model fit ( $\chi^2 / df = 2.58$ , CFI = .95, TLI = .95, RMSEA = .053, SRMR = .043). The scale demonstrated strong reliability (Cronbach's  $\alpha = .86$  overall; .93, .93, .88, .87, .91 for subscales). Item 20 is reverse-scored; total scores range from 22 to 110, with higher scores indicating stronger deficit-oriented attitudes. This context-specific scale offers a robust tool for investigating deficit-oriented beliefs in educational settings and provides a foundation for further research, teacher training, and policy development.

## 1. INTRODUCTION

Throughout history, the privileged structures of the dominant class have been legitimized through mechanisms that obscure reality. Within this context, the education system emerges as a structure that ostensibly allocates individuals to economic positions through an objective and seemingly meritocratic process (Bowles & Gintis, 2011). Contrary to these claims, however, education systems can significantly influence the outcomes that determine students' social positions. Educational institutions segregate students in ways that serve the interests of the dominant class, thereby shaping academic success within a social framework. As social class-based inequalities manifest as educational advantages and disadvantages in schools (Bourdieu & Passeron, 2015), millions of students from specific family backgrounds experience

\*CONTACT: İsmail ÇİMEN ✉ [ismailcimen@uludag.edu.tr](mailto:ismailcimen@uludag.edu.tr) 📍 Bursa Uludağ University, Faculty of Education, Department of Educational Sciences, Bursa, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

widespread and persistent academic failure (Valencia, 2010). In this regard, one of the key social determinants can be described as deficit thinking.

The origins of deficit thinking can be traced back to the colonization of Indigenous peoples, particularly by the British and other White populations, for economic gain. People of different racial and ethnic backgrounds were both disenfranchised through colonization and subjected to exploitation that was legitimized by attributing biological and cultural deficiencies to them. This ideology, used to justify the oppressor-oppressed relationship, has profoundly influenced the educational experiences of students from diverse racial and ethnic backgrounds (Menchaca, 1997). In this context, deficit thinking has evolved into one of the theories used to explain achievement gaps between advantaged and disadvantaged student groups. This mindset, which shapes education policies, attributes academic failures primarily to students and their families, largely ignoring the impact of societal and systemic factors on achievement (Valencia, 1997). Deficit theories assume that some children are inherently inferior to others due to genetic, cultural, or experiential differences and must overcome these deficiencies to learn effectively. Rather than examining the conditions provided to students, these theories place the blame for failure on perceived deficiencies within the students themselves (Nieto & Bode, 2018; Weiner, 2003). Consequently, proponents of deficit thinking fail to recognize the role of the existing education system in reinforcing social inequalities and shaping educational outcomes (Davis & Museus, 2019).

According to Valencia (1997), influenced by various theories such as Social Darwinism (Ornstein *et al.*, 2011; Russell, 1973), intelligence testing movements (Cronin *et al.*, 1974; Valencia & Suzuki, 2000) and “nature versus nurture” debates (Degler, 1991; Gillborn & Youdell, 2001; Payne, 2021), deficit thinking has six main themes. These include blaming the victim, oppression, pseudoscience, temporal changes, educability, and heterodoxy. Blaming the Victim, refers to the process in which a person who has been harmed by a crime is blamed for their victimization (The Canadian Resource Centre for Victims of Crime, 2022). Those who adhere to deficit thinking ignore the role of external and systemic factors in student success, instead claiming that academic failure is a result of cognitive or motivational deficiencies in the student (Valencia, 1997). Oppression, arises from power relations as a characteristic of social stratification and benefits the dominant culture (Marfin, 2019). Deficit thinking also represents a form of oppression, where power is used unfairly to maintain the position of a certain group (Gorski, 2010). Pseudoscience, refers to non-scientific practices based on ideological commitments or dogmatic thinking, yet it creates the illusion of science by using scientific language (Hyslop-Margison & Naseem, 2007; Thyer & Pignotti, 2016). Researchers who adopt deficit thinking also often violate scientific methods, attempting to rationalize deficit thinking with pseudoscientific data (Valencia, 1997). Temporal Changes, deficit thinking is shaped by the ideology of the time and the research climate, and the tools used to convey deficit thinking have undergone change over time (Valencia, 2010; Zammit, 2020). Educability, by attributing deficiencies or faults to students, deficit thinking particularly influences our views on the educability of students from lower socioeconomic backgrounds (Valencia, 1997). Although schools appear to have a mission to ensure every student’s learning, the actual practices and programs in schools reflect deficit thinking regarding the educability of children from low-income families (Scheurich & Skrla, 2004). Deficit thinking is based on orthodoxy, which refers to doctrines generally accepted and officially recognized within society. In contrast to the orthodox views that define deficit thinking, heterodox perspectives challenge this mindset and encourage the questioning of the status quo (Cormier, 2009). Heterodoxy, which opposes traditional norms and serves as a tool for adopting alternatives, plays a critical role in resisting deficit thinking by encouraging a re-evaluation of the conditions that foster and sustain this mindset (Huitt, 2023).

To combat the deficit thinking, which is rooted in systematic inequalities such as socioeconomic, racial, or ethnic differences (Sharma, 2009) and blames students for school



failure (Valencia, 2010), it is essential first to determine how this mindset manifests in teachers. Although several studies have investigated teachers' deficit thinking (Beasley, 2023; Cormier, 2009; Gray, 2015; Ireland, 2015; Lawrence, 2017; Marfin, 2019; Simone, 2012; Zammit, 2020), a significant limitation lies in the contextual specificity of their measurement tools. A review of the literature revealed that only one scale (Marfin, 2019) has been developed to directly measure deficit thinking. Other studies addressing deficit thinking have relied on related instruments, such as the Color-Blind Racial Attitudes Scale, Deficit-Oriented Questions, and the Attitude toward Poverty Scale (Accuardi-Gilliam, 2017; Austin, 2019; Gholson, 2015; Harper, 2010; Neville *et al.*, 2000; Simone, 2012; Yun & Weaver, 2010), which assess constructs that are conceptually aligned with or indicative of deficit-based perspectives. Moreover, the instruments used in these studies often include references that are culturally or racially grounded in the Western context, based on language, culture, race, ethnicity, and the Black-White dichotomy (e.g. students of color, Hispanic, Latino, African American, Black, or minority students)—that do not have meaningful equivalents in the Turkish context. For instance, Marfin's (2019, p. 156) questionnaire includes items such as "African American students and Hispanic students are more likely to be members of inappropriate groups than their White counterparts" or "In certain situations, I attribute a student's poor choices to the moral character of their ethnic or racial identity" or "Educators should be aware that students of color may not be able to reach a high level of academic achievement" which have no meaningful equivalents in Türkiye's demographic or cultural context. This cultural mismatch presents a major limitation for adapting these instruments directly to Türkiye. In contrast to Western countries, where ethnic or racial categories often serve as primary explanatory variables for educational disparities, in Türkiye, such differences are more commonly understood through a socioeconomic lens. That is, educational inequalities in Türkiye are predominantly attributed to students' socioeconomic background rather than to their racial or ethnic identity (Çiftçi & Çağlar, 2014; Education Reform Initiative (ERG), 2014; Ocak, 2024). Therefore, any attempt to measure deficit thinking in the Turkish context must focus primarily on attitudes related to class-based disadvantage, rather than ethnically coded attitudes. This contextual distinction is supported by a growing body of research. Studies conducted in Türkiye have consistently demonstrated a strong relationship between students' socioeconomic status and their academic performance, cognitive development, socio-emotional well-being, and long-term educational outcomes (ERG, 2014; Organization for Economic Co-operation and Development [OECD], 2018). Given these findings, measuring deficit thinking in Türkiye requires a context-specific instrument that accurately reflects the way teachers interpret and respond to class-based disparities in education. This study responds to that need by developing and validating a new scale that captures Turkish teachers' attitudes toward deficit thinking, grounded in both international theory and national sociocultural realities. It is hoped that this instrument will contribute not only to academic research but also to teacher training and policymaking efforts aimed at reducing educational inequalities. Specifically, this study was guided by the following two research questions (RQs):

RQ1. What are the underlying dimensions of Turkish teachers' deficit thinking as identified through empirical data?

RQ2. Can a valid and reliable measurement tool be developed to assess these dimensions within the Turkish educational context?

## 2. METHOD

This study, which involves a scale development process, was conducted using quantitative approaches. The research focuses on validity and reliability analyses, and the following steps were taken during the process in line with the scale development protocol proposed by DeVellis (2017). The present study followed a systematic, multi-step process. First, the construct to be measured—teachers' attitudes of deficit thinking—was clearly defined (Step 1). An extensive

literature review was conducted to generate a comprehensive item pool (Step 2), and Likert-type response options were selected as the measurement format (Step 3). The initial item pool was then reviewed by field experts specializing in social justice in education, educational inequality, and inclusive practices (Step 4). Informed by their feedback, items that were overly broad, redundant, or ambiguous were revised or removed. Although the inclusion of validation items (Step 5) was considered, the primary focus remained on the construct validity of the core scale. The revised scale was administered to a development sample (Step 6), and items were evaluated based on exploratory factor analysis (EFA) (Step 7). Items with low factor loadings and high cross-loadings were excluded during this phase. The final scale structure was optimized by removing non-performing items and ensuring conceptual clarity and parsimony (Step 8). Finally, the factor structure was evaluated through confirmatory factor analysis (CFA) to assess model fit and structural validity, in accordance with best practices for psychometric evaluation (Step 9). This rigorous protocol enhanced the construct validity and contextual relevance of the instrument. The details of the process have been explained below.

### 2.1. Definition of the Construct to be Measured

Deficit theories argue that student failure is attributed to deficiencies within the student, overlooking the underlying conditions that contribute to failure (Weiner, 2003). Teachers, who have an impact on student success, often continue to sustain deficit thinking towards disadvantaged students, even if it is not always the result of a conscious action (Gorski, 2010; Valencia, 2010). A review of the literature reveals that there is no measurement tool in Türkiye related to deficit thinking used to explain achievement gaps between student groups. As explained above, this study aims to fill this gap.

### 2.2. Item Pool Creation

According to Loevinger (1957), the items in the pool should be broad enough to encompass all alternative theories of the construct expected to be measured. Similarly, Clark and Watson (2016) emphasize the importance of systematically including all content relevant to the construct when developing an item pool. In their view, the initial item pool should be broad and comprehensive, and overinclusiveness should not be avoided at this stage. This is because psychometric analyses are strong in identifying irrelevant items that should be removed from the scale, but weak in identifying content that should be included (Boateng *et al.*, 2018). In line with this approach, based on the theoretical framework of deficit thinking, dimensions were identified, and a pool of 70 items was created by reviewing various studies. The items were formulated using a 5-point Likert scale, with response options ranging from "*Strongly Disagree* (1), *Disagree* (2), *Neutral* (3), *Agree* (4), to *Strongly Agree* (5)."

### 2.3. Expert Consultation

The draft form, consisting of 70 items, was first sent to a language expert for linguistic review, and the recommended revisions were implemented. Subsequently, to ensure content validity and gather feedback on the items, the form was reviewed by three academics with doctoral degrees specializing in social justice in education, educational inequalities, and inclusive education. The first academic identified items he found overly broad and recommended making them more specific. He also highlighted items with overlapping meanings and suggested their removal. Based on this feedback, the scope of the relevant items was narrowed, and redundant items were removed from the pool. The second academic emphasized that each item should express a single idea, drawing attention to the overuse of the conjunction "and." He also suggested replacing the phrase "is affected" with "is more affected" when comparing students from lower socioeconomic backgrounds with their peers (Final version of the scale: Items 12–16). In response, items were revised accordingly. The third academic offered suggestions to improve the clarity of two items (Final version: Items 13 and 18) and recommended the removal of one item. Revisions were made based on this feedback. The revised draft form was then resent to the three academics for further review. An online meeting followed, during which the

relevance, clarity, and cultural appropriateness of the items were evaluated, and consensus was reached. Finally, a researcher with a doctoral degree in measurement and evaluation reviewed the form. Based on this final review, the title of the scale was revised, and the finalized version of the scale consisted of 50 items.

## 2.4. Pilot Study

To assess the applicability and clarity of the developed form, a draft version of the scale was administered to twenty teachers selected through convenience sampling. Teachers were asked to read the items and identify any potential ambiguities or confusion in meaning. Based on the feedback received, several items were revised linguistically.

## 2.5. Data Collection

The final version of the draft scale was administered to a total of 352 teachers selected through convenience sampling from public schools located in the districts of Mudanya, Nilüfer, Osmangazi, and Yıldırım in Bursa during the 2023-2024 academic year. In the second phase, the scale was administered to 616 teachers using a cluster sampling method. For this stage, schools rather than individual teachers were sampled, and all teachers in the selected schools were invited to participate.

**Table 1.** Demographic information of the study group.

Categories	EFA		CFA	
	<i>f</i>	%	<i>f</i>	%
Gender				
Female	164	50.8	344	60.5
Male	159	49.2	225	39.5
Age				
20-30	17	5.3	33	5.8
31-40	146	45.2	162	28.5
41-50	124	38.4	250	43.9
51-60+	36	11.1	124	21.8
Professional Experience				
1-5	18	5.6	42	7.3
6-10	75	23.2	75	13.2
11-20	145	44.9	179	31.5
20+	85	26.3	273	48.0
Education Status				
Undergraduate	228	70.6	433	76.1
Master's Degree	92	28.5	131	23.0
PhD	3	.9	5	.9
SES				
Low	122	37.8	182	31.98
Middle	169	52.3	226	39.72
High	32	9.9	161	28.30
Total	323	100	569	100

SES: The socioeconomic status of the area where the school is located.

Data collection was conducted in coordination with school principals. Depending on the infrastructure and logistical preferences of the schools, the survey was either distributed via an

online link (Google Forms) or administered as a printed form. Participants completed the scale during designated periods that did not interfere with instructional hours. All participants were informed about the purpose of the research, the voluntary nature of participation, and the confidentiality of their responses through an introductory statement. No incentives were offered for participation. Data was screened for completeness before proceeding to analysis and only fully completed responses were included. Further details on the demographic characteristics of the participants—including variables such as gender distribution, teaching experience, and education status—are presented in Table 1.

## 2.6. Data Analysis

All statistical analyses were performed using IBM SPSS Statistics 29 and MPlus 8. Before the main analyses, the dataset was screened for missing values and outliers, and cases with incomplete responses or extreme values were removed. As a result, data from 29 participants in the first group and 47 participants in the second group were excluded (see Table 1).

Assumption checks were conducted prior to the main analyses. For the EFA sample ( $n = 323$ ), the adequacy of the sample size and the factorability of the correlation matrix were examined using the Kaiser–Meyer–Olkin (KMO) test and Bartlett’s Test of Sphericity. The KMO value was .872, indicating meritorious sampling adequacy (Kaiser, 1974). Bartlett’s Test of Sphericity yielded a chi-square value of 4983.993 ( $df = 253$ ,  $p < .001$ ), confirming that the correlation matrix was significantly different from an identity matrix and suitable for factor analysis (Can, 2016).

Multivariate normality was assessed with Mardia’s test. The results indicated violations of multivariate normality (Mardia Skewness = 30672.03, Kurtosis = 31.59,  $p < .001$ ). Consequently, Principal Axis Factoring (PAF) was selected as the extraction method. PAF does not require the assumption of multivariate normality and is therefore more appropriate for datasets where this assumption is violated (Fabrigar *et al.*, 1999; Osborne *et al.*, 2008). After deciding on the extraction method, in line with Thompson’s (2004) recommendation for research in the social sciences, Promax rotation was used as the preferred oblique rotation method. Promax is considered a more desirable choice when factors are expected to correlate, particularly in applied fields such as education and psychology (Thompson, 2004). Items with factor loadings below .40, communalities below .40, or cross-loadings with less than .10 difference were removed (Osborne *et al.*, 2008). The number of factors was determined based on eigenvalues  $> 1$ , the scree plot, and parallel analysis.

For the CFA, data from an independent sample ( $n = 569$ ) were used. Prior to CFA, multivariate normality was again examined using Mardia’s test, which indicated a violation of (Skewness = 5635.28,  $p < .001$ ; Kurtosis = 69.39,  $p < .001$ ), suggesting non-normality in the data distribution (Mardia, 1970). Given this result, the maximum likelihood robust method for CFA was employed in MPlus, which provides robust standard errors and chi-square statistics under non-normal conditions (Li, 2016). Additionally, the sample size ( $n = 569$ ) was deemed sufficient based on recommended criteria for CFA exceeding the commonly recommended threshold of 200 participants or a ratio of at least 10 participants per estimated parameter for CFA (Hair *et al.*, 2019; Kline, 2016). Linearity among variables was assumed, and factor correlation values did not indicate multicollinearity, as no correlation exceeded .85 (Brown, 2015).

After the assumption checks, EFA and CFA were conducted sequentially to evaluate construct validity. Model fit in CFA was evaluated using  $\chi^2/df$  ( $< 3$  for perfect fit), Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) ( $\geq .95$ ), Root Mean Square Error of Approximation (RMSEA) ( $\leq .08$ ), and Standardized Root Mean Square Residual (SRMR) ( $\leq .05$ ) as recommended by Hu and Bentler (1999) and Kline (2016). Following factor analyses, convergent and discriminant validity were examined by calculating Composite Reliability (CR), Average Variance Extracted (AVE), Maximum Shared Variance (MSV), and Average Shared Variance (ASV), and internal consistency reliability was assessed using Cronbach’s  $\alpha$ .

### 3. RESULTS

In this section, the results related to the analyses of validity and reliability are presented separately below.

#### 3.1. Exploratory Factor Analysis

While performing EFA using PAF with Promax rotation, Osborne *et al.* (2008) suggest that communalities above .40 are acceptable. Therefore, during the item retention process, items with factor loadings below .40 were excluded from the scale, as they were considered insufficiently representative of the underlying factors. Only Item 41's communality value was .38, but we chose to keep this item as it is important for theoretical considerations. Table 2 shows the communalities of the final version of the scale. In addition, to ensure distinct factor loadings, a minimum difference of .10 between an item's loadings on different factors was set as the threshold to identify and remove cross-loading items. As a result of these criteria, 27 items were excluded, leading to a scale composed of 23 items. However, in response to potential item redundancy within Factor 1, one item was excluded from the final scale. Subsequently, both the EFA and CFA were re-run with the revised 22-item scale. The results reported in this study are based on this final version. Further details regarding this decision can be found in the section for Reliability Analysis.

**Table 2.** Item communalities.

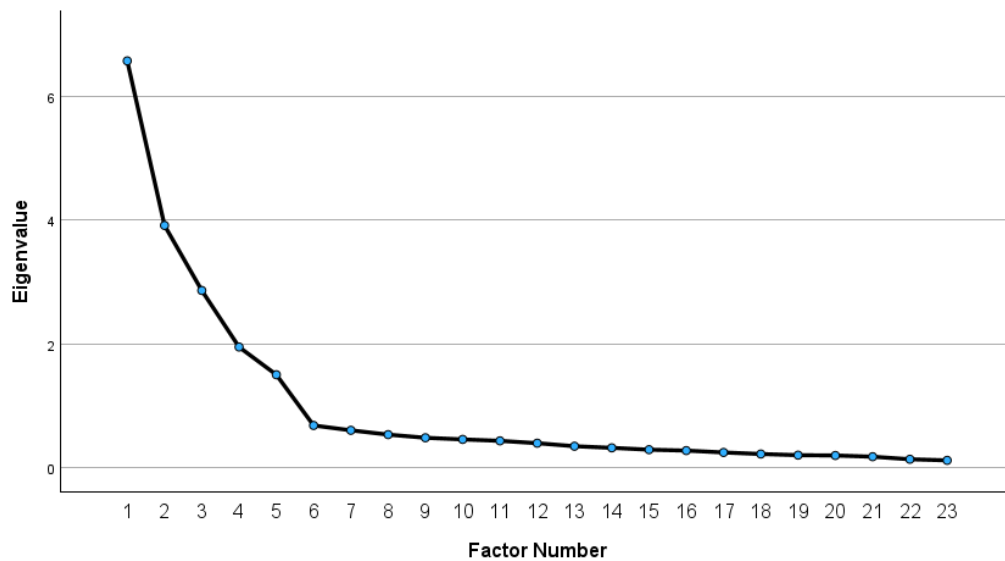
Item	Initial	Extraction	Item	Initial	Extraction
I1	.502	.499	I21	.786	.839
I2	.747	.868	I22	.820	.869
I3	.737	.813	I23	.762	.795
I11	.527	.511	I26	.486	.451
I12	.693	.694	I31	.532	.609
I13	.695	.732	I32	.574	.638
I15	.657	.696	I33	.623	.736
I16	.621	.656	I34	.410	.448
I17	.623	.607	I35	.625	.743
I18	.542	.534	I36	.611	.764
I20	.663	.668	I41	.360	.379

As shown in Table 2, all items included in the final version of the scale demonstrated acceptable levels of shared variance with the extracted factors. Specifically, communalities after extraction ranged from .379 to .869, indicating that a substantial proportion of each item's variance was explained by the factor structure. Although one item (I41) had a communality slightly below the commonly accepted threshold of .40 (Osborne *et al.*, 2008), it was retained due to its strong theoretical relevance to the construct. All other items met or exceeded the recommended .40 criterion, supporting the adequacy of their inclusion in the final factor model.

The number of factors to retain was determined based on multiple criteria: eigenvalues greater than 1.0, inspection of the scree plot, and the results of a parallel analysis. These criteria collectively suggested a five-factor solution, which was also theoretically meaningful. To determine the number of factors to retain, we adopted a multi-step procedure combining empirical criteria with theoretical judgment. Following the protocol by Yurt (2024), initially, we reviewed communalities and factor loadings to assess the suitability of each item. Items with low communalities or insufficient factor loadings were removed. After each round of item removal, the factor analysis was re-run to evaluate the updated structure. Following this



iterative process, a parallel analysis was conducted, the results of which, supported by a scree plot (See [Figure 1](#)), indicated a five-factor solution as the most appropriate.



**Figure 1.** Scree plot for eigenvalues.

Upon examining the distances between each point representing the factors and the inflection points of the lines in the line graph in [Figure 1](#), it can be concluded that a five-factor structure emerges. This finding is also consistent with the factor eigenvalues presented in [Table 3](#). A final EFA was conducted using this predefined factor structure. During this stage, additional items were removed based on statistical and conceptual grounds. The final factor structure was reviewed and confirmed to be theoretically sound and empirically supported.

**Table 3.** Results of parallel analysis.

Real Data		Random Data	
Factor	Eigenvalue	Factor	Eigenvalue
1	8.284	1	1.626
2	4.781	2	1.539
3	3.589	3	1.470
4	2.105	4	1.417
5	1.814	5	1.369
6	1.255	6	1.325
7	1.021	7	1.284

One of the methods used to determine the number of factors in the data is parallel analysis, which operates based on randomly generated data. The core of the analysis involves comparing the eigenvalues of the randomly generated data with the eigenvalues obtained from the real data set. The point where the eigenvalue of the parallel data exceeds the eigenvalue of the real data provides information about the number of factors (Ledesma & Mora, 2007). According to this, when [Table 3](#) is examined, it is observed that after the fifth dimension, the values in the parallel data exceed those in the real data. Therefore, the five-factor structure of the scale is also confirmed by the parallel analysis method. According to [Table 4](#), the scale consists of 5 factors, which account for % 66.13 of the total variance of the scale. After rotation, it was observed that the first factor consists of 7 items, the second factor consists of 5 items, the third factor consists of 4 items, and the fourth and fifth factors each consist of 3 items. In alignment with the definitions in the literature, the first factor was named "Blaming the Environment," the second

factor "Educability," the third factor "Oppression," the fourth factor "Blaming the Victim," and the fifth factor "Pseudoscience".

**Table 4.** Pattern matrix and explained variances.

Items	The factor loading values after rotation				
	1	2	3	4	5
I13	<b>.874</b>	-.057	.023	-.031	-.016
I12	<b>.862</b>	.033	-.017	-.059	.002
I15	<b>.846</b>	-.012	.090	-.049	-.053
I17	<b>.816</b>	-.037	-.005	.025	.016
I16	<b>.729</b>	-.009	-.027	.070	.058
I11	<b>.726</b>	.052	-.008	.004	.005
I18	<b>.662</b>	-.057	.023	-.031	-.016
I22	.008	<b>.934</b>	-.023	.005	.018
I21	-.031	<b>.916</b>	-.019	-.014	.001
I23	.003	<b>.890</b>	-.002	-.009	-.019
I20	.024	<b>.814</b>	-.017	.051	.003
I26	-.008	<b>.656</b>	.094	-.064	-.013
I33	.089	.023	<b>.859</b>	.012	.035
I32	.004	.038	<b>.802</b>	.060	.000
I31	-.010	.052	<b>.779</b>	.024	-.007
I34	.100	.108	<b>-.621</b>	.093	.044
I2	-.031	.010	.016	<b>.951</b>	-.045
I3	-.001	-.028	.013	<b>.902</b>	.023
I1	.152	-.012	-.012	<b>.609</b>	.018
I36	-.032	.026	-.017	.011	<b>.880</b>
I35	-.075	-.033	.075	-.032	<b>.843</b>
I41	.145	-.005	-.067	.006	<b>.592</b>
Explained Variance	%25.26	%16.65	%11.41	%7.28	%5.18
(Total: 66.13%)					

The initial item pool was informed by the six conceptual dimensions of deficit thinking described by Valencia (1997, 2010). Although we theoretically anticipated a multidimensional structure aligned with these dimensions, the EFA revealed a five-factor solution. Items related to temporal changes and heterodoxy did not demonstrate sufficient loadings and were excluded. Instead, a new factor, Blaming the Environment, emerged, which is consistent with environmental deficit perspectives discussed in prior research (Pearl, 1997; Valencia, 2020)

### 3.2. Confirmatory Factor Analysis

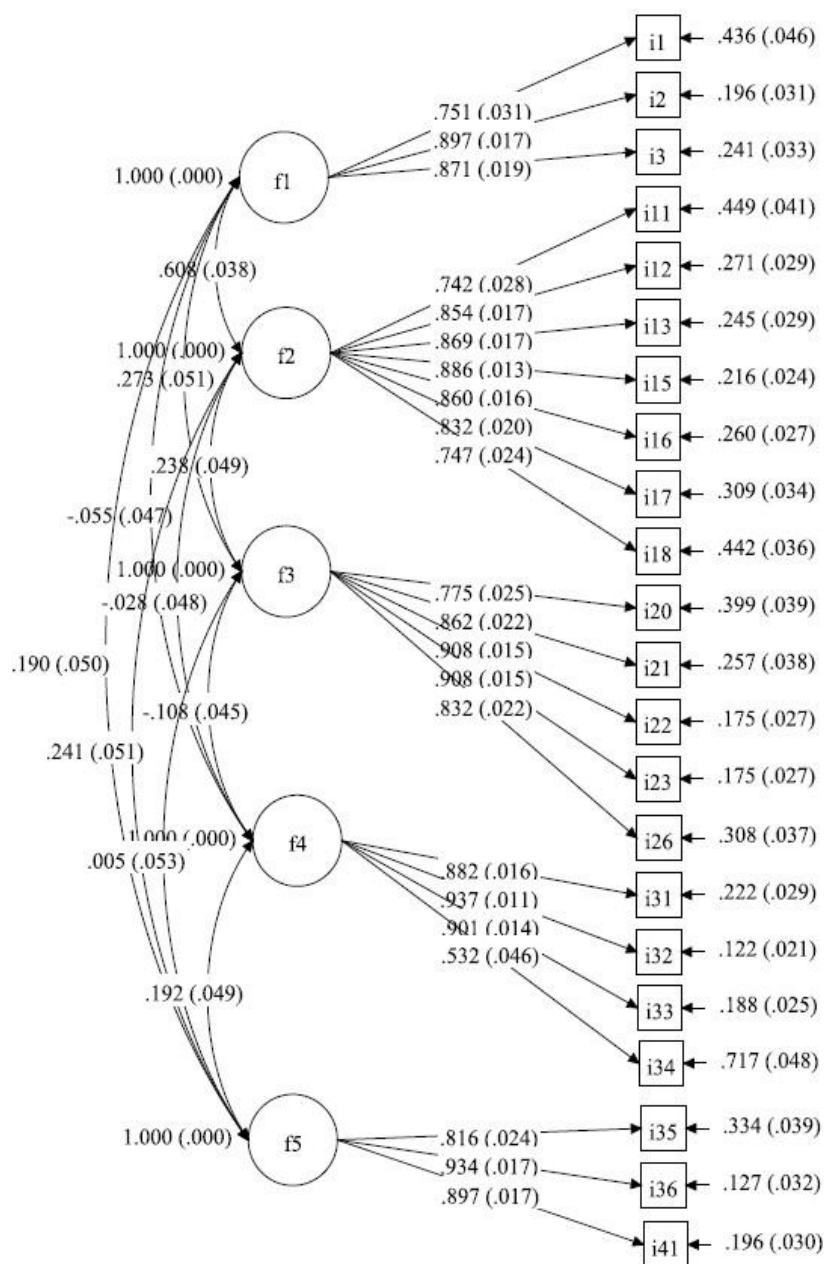
After EFA, the 22-item form of the scale was administered to 569 individuals for CFA to evaluate the structure of the 5-factors. Table 5 presents the fit indices of the model along with their respective value ranges. According to Table 5 the model demonstrated an overall good fit to the data:  $\chi^2/df = 2.58$ , CFI = .95, TLI = .95, RMSEA = .053, and SRMR = .043. The model fit was evaluated based on multiple fit indices following the recommended cut-off values by Hu and Bentler (1999), Hooper *et al.* (2008), and Kline (2016). According to established criteria values of CFI and TLI above .95 and SRMR below .05 indicate an excellent fit, while an RMSEA value between .05 and .08 suggests acceptable fit. These results support the adequacy

of the hypothesized factor structure in representing the data. Furthermore, the Akaike Information Criterion (AIC = 25840.88), the Bayesian Information Criterion (BIC = 26171.01), and adjusted BIC (25929.75) were reported to allow model comparison in future studies.

**Table 5.** CFA Results.

Fit Index	Perfect Fit Index	Acceptable Fit Index	Values for the Scale	Fit Result
$\chi^2 / df$	< 3	< 5	2.58	Perfect
CFI	.95 ≤ CFI ≤ 1.00	.90 ≤ CFI ≤ .95	.95	Perfect
TLI	.95 ≤ TLI ≤ 1.00	.90 ≤ TLI ≤ .95	.95	Perfect
RMSEA	.00 ≤ RMSEA ≤ .05	.05 ≤ RMSEA ≤ .08	.053	Acceptable
SRMR	.00 ≤ SRMR ≤ .05	.05 ≤ SRMR ≤ .10	.043	Perfect

$\chi^2/df$ : Chi-Square / Degrees of Freedom.



**Figure 2.** Standardized path diagram of the CFA.

When Figure 2 is examined, it is observed that the factor loadings range from .53 to .94. These values indicate that the model is a good fit. Detailed information is provided below in Table 6.

**Table 6.** Standardized factor loadings, standard errors and critical ratios.

Item	Standardized Loading	SE	Critical Ratio	<i>p</i>	Item	Standardized Loading	SE	Critical Ratio	<i>p</i>
I1	.751	.031	24.58	.000	I21	.862	.022	38.64	.000
I2	.897	.017	52.38	.000	I22	.908	.015	61.37	.000
I3	.871	.019	46.30	.000	I23	.908	.015	61.06	.000
I11	.742	.028	26.82	.000	I26	.832	.022	37.44	.000
I12	.854	.017	50.65	.000	I31	.882	.016	53.46	.000
I13	.869	.017	52.30	.000	I32	.937	.011	83.85	.000
I15	.886	.013	66.28	.000	I33	.901	.014	63.74	.000
I16	.860	.016	55.39	.000	I34	.532	.046	11.69	.000
I17	.832	.020	41.01	.000	I35	.816	.024	34.21	.000
I18	.747	.024	30.92	.000	I36	.934	.017	54.52	.000
I20	.775	.025	30.55	.000	I41	.897	.017	54.28	.000

While evaluating the standardized factor loadings, .40 was applied as a minimum acceptable loading to ensure that each item contributed meaningfully to its latent construct in line with (Hair *et al.*, 2019; Osborne *et al.*, 2008). All items loaded significantly on their respective latent factors, with standardized factor loadings ranging from .532 to .937. All loadings were statistically significant at the  $p < .001$  level. While I34, associated with Factor 4, exhibited the lowest loading value (.532), it remained statistically significant and did not substantially impair the overall model fit. Critical ratio values, which represent the parameter estimate divided by its standard error, ranged from 11.69 to 83.85, further confirming the precision and significance of the estimated loadings (Hair *et al.*, 2019). These findings indicate that the items demonstrate a strong and coherent representation of their respective constructs, supporting the internal consistency of each factor.

### 3.3. Convergent and Discriminant Validity

To evaluate the discriminant validity of the factors extracted, the factor correlation matrix was examined (see Table 7). As the analysis employed Promax rotation, which allows factors to correlate, the observed inter-factor correlations ranged from -.147 to .548. While moderate correlations (e.g., between Factor 1 and Factor 4) suggest conceptual relatedness, the overall low-to-moderate correlation values provide evidence that the factors are sufficiently distinct. These results support the notion that the identified factors represent separate dimensions of deficit thinking, which is consistent with the theoretical framework underpinning the scale.

**Table 7.** Factor correlation matrix.

Factors	1	2	3	4	5
F1	1				
F2	-.005	1			
F3	-.147	.043	1		
F4	.548	.016	-.135	1	
F5	.066	-.089	.160	.050	1

In CFA models, convergent validity refers to the relationship between the items that form the scale and their corresponding factor. Discriminant validity, on the other hand, refers to the

relationship between the items and their factor being stronger than their relationships with other factors (Gürbüz, 2021). The fact that subdimensions come together to form a structure depends on a certain level of correlation between them. Conversely, the ability of each dimension to form is dependent on the dimensions being distinct from each other, i.e., their separation. This necessitates the calculation of convergent and discriminant validity (Bülül & Demirer, 2008). For this purpose, the CR, AVE, MSV, and ASV values for the factors of the scale are presented. Specifically, AVE and CR values were calculated following the formulas recommended by Gürbüz (2021). AVE was computed by averaging the squared standardized factor loadings of the items associated with each latent construct. CR was calculated by summing the standardized loadings, squaring the sum, and dividing it by the squared sum plus the sum of error variances. The MSV and ASV values were obtained using inter-factor correlations. MSV represents the square of the highest correlation of a construct with any other construct in the model. ASV was calculated as the average of all squared correlations of a given construct with the remaining constructs. These values were used to assess discriminant validity by comparing them with the corresponding AVE values of Gürbüz (2021).

Upon examining Table 8, it can be observed that all factors exhibit high reliability ( $CR > .7$ ). The AVE values for the factors, being above .5 and lower than the CR values, indicate convergent validity. Furthermore, the MSV and ASV values of the factors, being lower than the AVE values, confirm discriminant validity (Gürbüz, 2021).

**Table 8.** Convergent and discriminant scores.

Factors	CR	AVE	MSV	ASV
F1	.88	.71	.37	.12
F2	.94	.69	.37	.12
F3	.93	.74	.08	.04
F4	.89	.69	.04	.01
F5	.91	.78	.06	.03

### 3.4. Reliability Analysis

In addition to the CR coefficient provided in Table 8, Cronbach's Alpha coefficient has been calculated to determine the reliability of the scale and subdimensions. Moreover, McDonald's Omega ( $\omega$ ) was also computed to provide a more robust estimate of internal consistency for the full scale ( $\omega = .874$ ). The reliability values for the scale and its factors, which were applied to 569 teachers for the CFA analysis, are presented in Table 9.

**Table 9.** Cronbach's alpha coefficients.

Factors	$\alpha$
Total (22 items)	.86
Factor 1 (7 items)	.93
Factor 2 (5 items)	.93
Factor 3 (4 items)	.88
Factor 4 (3 items)	.87
Factor 5 (3 items)	.91

Cronbach's Alpha values above .70 are generally considered acceptable, while values above .80 indicate good reliability (Kline, 2016). Therefore, the total scale and its factors have good reliability. The McDonald's Omega ( $\omega$ ) value for the full scale was .874, further supporting the internal consistency of the instrument.



#### 4. DISCUSSION and CONCLUSION

The purpose of the study was to develop a valid and reliable scale to determine teachers' deficit thinking. To achieve this goal, a literature review was first conducted, and the content of deficit thinking was explored. This led to the creation of a draft form consisting of 70 items. After expert consultations, the number of items in the draft form was reduced to 50. The 50-item draft form was administered to 352 teachers, followed by EFA. After EFA, the 5-dimensional, 22-item form was administered to 616 teachers. Data from 76 participants were removed from the dataset after outlier and missing data checks. Ultimately, validity and reliability studies were conducted on data from 892 teachers. After the analyses, the "Deficit Thinking Scale for Teachers" was developed, which is both valid and reliable for teachers.

The 22-item Deficit Thinking Scale for Teachers is a 5-factor scale with a 5-point Likert type scale ([Appendix 1](#)). Item 20 in the scale should be reverse-coded. The Cronbach's Alpha reliability coefficient for the entire scale is .86, and for the subdimensions, it is .93 for the first factor, .93 for the second factor, .88 for the third factor, .87 for the fourth factor, and .91 for the fifth factor. The lowest possible score on the scale is 22, and the highest possible score is 110. As the score on the scale increases, it is assumed that the teachers' deficit thinking also increases.

The factors of the scale are named as Blaming the Environment, Educability, Oppression, Blaming the Victim, and Pseudoscience. For the Blaming the Environment factor, representative items include statements such as "Families of low socioeconomic status participate less in parent-teacher meetings than other families." which reflect how failure is attributed to family-related variables. In this dimension, it is claimed that families are blamed for students' academic failure due to their inadequacies or dysfunctions. The Educability factor includes items like "The academic achievement of students from low socioeconomic backgrounds is more affected by the teacher's individual attention than that of other students" or "The academic achievement of students of low socioeconomic backgrounds is more affected by the teacher's feedback than that of other students" emphasizing perceived limitations in students' learning potential. In this dimension, it is argued that students' failure is independent of the conditions in society and schools, and if no intervention is made, this failure will continue. The Oppression factor is represented by items such as "The education system provides the necessary learning conditions for students from low socioeconomic backgrounds" or "The education system provides opportunities for students from low socioeconomic backgrounds to succeed" reflecting the so-called meritocratic discourse that responsibility is fulfilled at the system level. This dimension is based on the idea that the education system, particularly for students with certain backgrounds, serves to condemn them to poverty, thus controlling the flow of social status. The Blaming the Victim factor features statements such as "Students from low socioeconomic backgrounds are more likely to exhibit aggressive behavior than other students" or "Students from low socioeconomic backgrounds are more likely to exhibit disruptive behavior than other students" attributing failure directly to student characteristics. This dimension is based on finding fault in those who are victims of inequality, justifying the inequality, and in this dimension, students are blamed for educational outcomes. Lastly, the Pseudoscience factor includes items like "The academic failure of students from low socioeconomic backgrounds stems from their failure to make the necessary effort" or "The academic failure of students from low socioeconomic backgrounds stems from their lack of motivation to learn" capturing the tendency to explain failure through individual traits while ignoring structural factors. In this dimension, academic failure is attributed to a lack of students' abilities, intelligence, or effort, without considering the social structure as a variable affecting educational outcomes.

Although the theoretical framework of deficit thinking includes six core dimensions as identified by Valencia (1997, 2010)—namely blaming the victim, oppression, pseudoscience,

temporal changes, educability, and heterodoxy—the final structure of the “Deficit Thinking Scale for Teachers” consists of five empirically derived factors. During the EFA, items related to the dimensions of *temporal changes* and *heterodoxy* did not demonstrate sufficient factor loadings or distinctiveness and were therefore removed to ensure construct validity. This outcome may reflect the more abstract and philosophical nature of these two dimensions, which can be conceptually rich but difficult to translate into attitudinal indicators for a Likert-scale instrument. For example, the elimination of the notion of deficit based on orthodox views requires the adoption of heterodox perspectives or the questioning of what is considered unquestionable. According to Bourdieu (1977), such transformation typically occurs through a debate or crisis within a class society, which necessitates a fundamental re-evaluation of the structure of mindset (Huitt, 2023). However, it may often be difficult to observe or measure such ideological shifts through standardized items. Similarly, *temporal changes* refer to the variation in the vehicles through which deficit thinking is transmitted, depending on the spirit and conditions of the time (Valencia, 2010). For instance, while students' deficits were attributed to genetics, they were later linked to culture—showcasing the historically changeable nature of deficit thinking. Yet, this dynamic and diachronic aspect may not manifest clearly in present-day teacher attitudes. Additionally, the findings of the present study are also consistent with those of Marfin (2019), who developed a similar Deficit Thinking Questionnaire. In Marfin's study as well, the dimensions of heterodoxy and temporal changes were not included in the final structure. Their exclusion in both studies reinforces the notion that certain abstract or philosophical aspects of deficit thinking may not readily translate into measurable attitudinal indicators. In contrast, while the original theoretical model did not include *Blaming the Environment* as a separate dimension, its empirical emergence as an independent factor is both theoretically and sociologically meaningful. Although *blaming the victim* is the most prominent feature that captures the core logic of the deficit mindset and illustrates how this ideology is enacted in practice, the deficit perspective extends beyond the individual. Those who adopt this mindset also direct blame toward the victim's family, culture, and broader environment (Valencia, 2010). Notably, during the 1960s, an *accumulated environmental deficit model* was developed to explain school failure, particularly among poor children and minority groups, attributing their academic underachievement to accumulated deficits in their home and family environments (Pearl, 1997; Valencia, 2020). Furthermore, this finding is consistent with localized sociocultural dynamics and the existing literature on teacher discourse in the context of Türkiye, where student underachievement is frequently attributed to parental negligence or lack of involvement (Aydoğdu-Bilgiç, 2024; Dinçer, 2015; Orman, 2012).

#### 4.1. Limitations

This study had several limitations. First, convenience sampling was used in the first phase in order to collect data quickly and easily, especially considering that the large number of items in the draft form required voluntary participation from teachers. Second, as this study was designed to measure attitudes toward deficit thinking, it may involve limitations specific to attitude research. Although attitudes can be predictors of behavior, it should be acknowledged that they may be influenced by social pressure (Shrigley, 1983) or participants may provide responses that are socially desirable rather than ones that genuinely reflect their personal attitudes (Callegaro, 2008; DeVellis, 2017). Third, the reliability estimates for some factors exceeded .90, which may suggest potential item redundancy (Streiner, 2003). Therefore, future research may consider revisiting those factors for possible scale refinement.

#### 4.2. Practical Implications

School administrators can use this study to evaluate teachers' attitudes toward deficit thinking and identify their schools' professional development needs accordingly. The quantitative data obtained from the scale can serve as an evidence-based resource for developing policies focused on social justice in education. Additionally, the scale can be employed within undergraduate

teacher education programs to monitor pre-service teachers' attitudes toward deficit-oriented thinking. By administering the scale at the beginning and end of the academic year, it becomes possible to track changes in these attitudes over time.

#### 4.3. Recommendations for Future Research

Future research could establish a conceptual framework for uncovering the sociocultural and psychological structures associated with deficit thinking. In this context, it is essential to investigate the antecedents that give rise to deficit-oriented perspectives, as well as the consequences such frameworks produce within institutional, educational, and interpersonal domains. Moreover, both the direct and mediated effects of deficit thinking should be subjected to rigorous empirical investigation, enabling a more nuanced understanding of its role in reproducing structural inequalities and shaping individual-level outcomes. This study employed a quantitative research method. In order to gain deeper insights into how deficit thinking manifests within educational settings in Türkiye, qualitative research methods should also be utilized in future studies to complement quantitative findings. This research was conducted in the central districts of Bursa. To identify potential regional differences in deficit thinking and to enable more comprehensive interpretations at the national level, it is recommended that similar studies be carried out in other regions of Türkiye, involving a broader and more diverse sample of teachers. This study developed a scale to measure teachers' attitudes toward students' attributed deficits. However, as emphasized in the literature, individuals who are subjected to domination may unknowingly contribute to or even consent to their own domination (Bourdieu, 2019; Gramsci, 1999). In this regard, developing scales to assess students' attitudes toward their own attributed deficits, as well as parents' attitudes toward their children's attributed deficits, could offer a more comprehensive understanding of the multidimensional nature of deficit thinking. Such studies would provide deeper insights into the role of students and parents in the reproduction of this mindset. Despite its limitations, the items developed in this study may serve as a valuable foundation for future research on deficit thinking in educational contexts.

#### Acknowledgments

This study was derived from the master's thesis of the first author, conducted under the supervision of the second author. Additionally, it was presented at the 17th International Congress on Educational Administration, İstanbul Medeniyet University, on March 7–9, 2024, in İstanbul.

#### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number:** Bursa Uludağ University, 01.26.2024 / 2024-01 / Decision Number 32.

#### Contribution of Authors

**Abide Ocak:** Writing-original draft, Investigation, Resources, Methodology, Analysis, Validation, and Visualization. **İsmail Çimen:** Resources, Software, Analysis, Critical review, Methodology.

#### Orcid

Abide Ocak  <https://orcid.org/0009-0006-0274-6983>

İsmail Çimen  <https://orcid.org/0000-0001-9160-2856>

#### REFERENCES

Accuardi-Gilliam, J.E. (2017). *Examining the gap: Teachers' color-blind racial ideology and deficit thinking through the lens of school discipline* [Doctoral dissertation, Lewis and Clark College]. ProQuest Dissertations & Theses Global.

- Austin, C. (2019). *A grounded theory survey study of teachers' perception perpetuating the deficit narrative about marginalized students of color* [Master's thesis, University of Central Florida]. ProQuest Dissertations & Theses Global.
- Aydoğdu-Bilgiç, M. (2024). *Roman öğrencilerin okula devamsızlık, okul terki ve okuldaki başarısızlıklarına ilişkin öğrenci, öğretmen ve veli görüşleri* [Opinions of students, teachers, and parents regarding Roman students' absenteeism, dropping out of school, and academic failure] [Unpublished master's thesis]. Sakarya University.
- Beasley, C. (2023). *The miseducation of the black teacher: An examination of anti-blackness and deficit thinking* [Doctoral dissertation, Wayne State University]. ProQuest Dissertations & Theses Global.
- Boateng G.O., Neilands T.B., Frongillo E.A., Melgar-Quinonez H.R., & Young S.L. (2018) Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health* 6, Article 149. <http://doi.org/10.3389/fpubh.2018.00149>
- Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge University Press.
- Bourdieu, P., & Passeron, J.C. (2015). *Yeniden üretim: Eğitim sistemine ilişkin bir teorinin ilkeleri* [Reproduction: Principles of a theory of the education system] (A. Sümer, L. Ünsaldı & Ö. Akkaya, Trans.). Heretik Publications. (Original work published 1970)
- Bourdieu, P. (2019). *Eril tahakküm* [Male domination] (5th ed.) (B. Yılmaz, Trans.). Bağlam Publishing. (Original work published 1998)
- Bowles, S., & Gintis, H. (2011). *Schooling in capitalist America: Educational reform and the contradictions of economic life*. Haymarket Books.
- Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. Guilford Press.
- Bülbül, H., & Demirel, Ö. (2008). Hizmet kalitesi ölçüm modelleri Servqual ve Serperf'in karşılaştırmalı analizi [Comparative analysis of service quality measurement models Servqual and Serperf]. *Selçuk University Journal of Institute of Social Sciences*, (20), 181-198.
- Callergaro, M. (2008). Social desirability. In P.J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (Vol. 2, pp. 824-825). Sage Publications.
- Can, A. (2016). *SPSS ile bilimsel araştırma sürecinde nicel veri analizi* [Quantitative data analysis in the process of scientific research with SPSS]. Pegem Academy.
- Clark, L.A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A.E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (4<sup>th</sup> ed., pp. 187-203). American Psychological Association. <https://doi.org/10.1037/14805-012>
- Cormier, B.D. (2009). *Deconstructing the deficit-thinking paradigm in district and campus level leadership to close the achievement gap* [Doctoral dissertation, The University of Texas at Austin]. ProQuest Dissertations & Theses Global.
- Cronin, J., Daniels, N., Hurley, A., Kroch, A., & Webber, R. (1974). Race, class, and intelligence: A critical look at the IQ controversy. *International Journal of Mental Health*, 3(4), 46-132. <https://www.jstor.org/stable/41344019>
- Çiftçi, C., & Çağlar, A. (2014). Ailelerin sosyoekonomik özelliklerinin öğrenci başarısı üzerindeki etkisi: Fakirlik kader midir? [The impact of families' socioeconomic characteristics on student achievement: Is poverty destiny?]. *International Journal of Human Sciences*, 11(2), 155-175. <https://doi.org/10.14687/ijhs.v11i2.2914>
- Davis, P., & Museus, S. (2019). What is deficit thinking? An analysis of conceptualizations of deficit thinking and implications for scholarly research. *Currents*, 1(1), 117-130. <http://dx.doi.org/10.3998/currents.17387731.0001.110>
- Degler, C.N. (1991). *In search of human nature: The decline and revival of Darwinism in American social thought*. Oxford University Press.
- DeVellis, R.F. (2017). *Scale development: Theory and applications*. SAGE Publications.



- Dinçer, M. (2015). *A study on the dynamics of parent satisfaction and student academic achievement at schools using system dynamics modeling* [Unpublished doctoral dissertation]. Yeditepe University.
- Education Reform Initiative. (2014, May). *Equity and academic achievement in the Turkish education system: Research report and analysis*. Sabancı University.
- Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299. <https://doi.org/10.1037/1082-989X.4.3.272>
- Gholson, M.L. (2015). *Rural principal attitudes toward poverty and the poor* [Doctoral dissertation, Ohio University]. ProQuest Dissertations & Theses Global.
- Gillborn, D., & Youdell, D. (2001). The new IQism: Intelligence, 'ability' and the rationing. In J. Demaine (Ed.), *Sociology of education today* (pp. 65-97). Palgrave. [https://doi.org/10.1057/9780333977507\\_5](https://doi.org/10.1057/9780333977507_5)
- Gorski, P.C. (2010). Unlearning deficit ideology and the scornful gaze: Thoughts on authenticating the class discourse in education. *Counterpoints*, 402, 152-173. <https://www.jstor.org/stable/42981081>
- Gramsci, A. (1999). *Selections from the prison notebooks*. Elecbook.
- Gray, T.T. (2015). *Educator perceptions of deficit thinking and deficit thinking's influence on student achievement in secondary urban schools in North Texas* [Doctoral dissertation, Texas Wesleyan University]. ProQuest Dissertations & Theses Global.
- Gürbüz, S. (2021). *Amos ile yapısal eşitlik modellemesi* [Structural equation modeling with Amos]. Seçkin Publishing.
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2019). *Multivariate data analysis*. Cengage Learning.
- Harper, S.R. (2010). An anti-deficit achievement framework for research on students of color in STEM. *New Directions for Institutional Research*, 2010(148), 63-74. <https://doi.org/10.1002/ir.362>
- Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, 6(1), 53-60.
- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huitt, F.C. (2023). *Challenging deficit-thinking views about English language learners* [Doctoral dissertation, The University of Arizona]. ProQuest Dissertations & Theses Global.
- Hyslop-Margison, E.J., & Naseem, M.A. (2007). *Scientism and education: Empirical research as neo-liberal ideology*. Springer Science & Business Media.
- Ireland, J.H.L. (2015). *Disrupting deficit ideologies that impact learning for English language learners: An elementary principal's role* [Doctoral dissertation, Washington State University]. ProQuest Dissertations & Theses Global.
- Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31-36. <https://doi.org/10.1007/BF02291575>
- Kline, R.B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.
- Lawrence, S.M. (2017). *The impact of an action research study on deficit thinking: In an elementary school* [Doctoral dissertation, The University of North Carolina at Greensboro]. ProQuest Dissertations & Theses Global.
- Ledesma, D.R., & Mora, P.V. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation*, 12(1), Article 2. <https://doi.org/10.7275/wjnc-nm63>
- Li, C.H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, 21(3), 369-387. <https://doi.org/10.1037/met0000093>



- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635-694.
- Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519-530. <https://doi.org/10.1093/biomet/57.3.519>
- Marfin, B.L. (2019). *The role of deficit thinking in educational outcomes: Developing a questionnaire to determine adherence to deficit thinking* [Doctoral dissertation, Prairie View A&M University]. ProQuest Dissertations & Theses Global.
- Menchaca, M. (1997). Early racist discourses: The roots of deficit thinking. In R.R. Valencia (Ed.), *The evolution of deficit thinking* (pp. 13-40). Routledge.
- Neville, H.A., Lilly, R.L., Duran, G., Lee, R.M., & Browne, L. (2000). Construction and initial validation of the color-blind racial attitudes scale (CoBRAS). *Journal of Counseling Psychology*, 47(1), <https://doi.org/10.1037/0022-0167.47.1.59>
- Nieto, S., & Bode, P. (2018). *Affirming diversity: The sociopolitical context of multicultural education* (7<sup>th</sup> ed.). Pearson Education.
- Ocak, A. (2024). *Okul yöneticilerinin sosyal adalet liderliğinin öğretmenlerin eksiklik düşüncesindeki rolü* [The role of school principals' social justice leadership in teachers' deficit thinking] [Unpublished master's thesis]. Uludağ University.
- OECD. (2018). *Equity in education: Breaking down barriers to social mobility, PISA*. OECD Publishing. <https://doi.org/10.1787/9789264073234-en>
- Orman, M. (2012). *Velilerin sosyo ekonomik durumu, sınıf veli toplantılarına katılımı ve öğrencilerin başarıları (Tahir Merzeci ilköğretim okulu örneği)* [Parents' socioeconomic status, participation in class parent meetings, and student achievement (The case of Tahir Merzeci primary school in İzmir)] [Unpublished master's thesis]. Dokuz Eylül University.
- Ornstein, A., Levine, D., Gutek, G., & Vocke, D. (2011). *Foundations of education*. Wadsworth, Cengage Learning.
- Osborne, J., Costello, A., & Kellow, J. (2008). Best practices in exploratory factor analysis. In J. Osborne (Ed.) *Best practices in exploratory factor analysis* (pp. 86-99). SAGE Publications.
- Payne, J.M. (2021). Rethinking nature and nurture in education. *Journal of Philosophy of Education*, 55(1), 143-166. <https://doi.org/10.1111/1467-9752.12527>
- Pearl, A. (1997). Cultural and accumulated environmental deficit models. In R.R. Valencia (Ed.), *The evolution of deficit thinking* (pp. 132-159). Routledge.
- Russell, B. (1973). *Batı felsefesi tarihi 3 (Modern çağ) (Yeni çağ)* [History of Western Philosophy 3 (Modern Age) (New Age)] (M. Sencer, Trans.). Bilgi Publications. (Original work published 1945)
- Scheurich, J.J., & Skrla, L. (2004). *Educational equity and accountability: Paradigms, policies and politics*. Psychology Press.
- Sharma, M. (2009). *Inner city students: Stamped, labeled and shipped out! Deficit thinking and democracy in an age of neoliberalism* [Doctoral dissertation, University of Toronto]. ProQuest Dissertations & Theses Global.
- Shrigley, R.L. (1983). The attitude concept and science teaching. *Science Education*, 67(4), 425-442. <https://doi.org/10.1002/sce.3730670402>
- Simone, J.A. (2012). *Addressing the marginalized student: The secondary principal's role in eliminating deficit thinking* [Doctoral dissertation, University of Illinois at Urbana-Champaign]. ProQuest Dissertations & Theses Global.
- Streiner, D.L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, 80(1), 99-103. [https://doi.org/10.1207/S15327752JPA8001\\_18](https://doi.org/10.1207/S15327752JPA8001_18)
- The Canadian Resource Centre for Victims of Crime. (2022). <https://crcvc.ca>
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

- Thyer, B.A., & Pignotti, M. (2016). The problem of pseudoscience in social work. *Journal of Social Work Education*, 52(2), 136-146. <https://doi.org/10.1080/10437797.2016.1151279>
- Valencia, R.R. (1997). *The evolution of deficit thinking: Educational thought and practice*. The Falmer Press.
- Valencia, R.R. (2010). *Dismantling contemporary deficit thinking: Educational thought and practice*. Routledge.
- Valencia, R.R., & Suzuki, L.A. (2000). *Intelligence testing and minority students: Foundations, performance factors, and assessment issues*. Sage Publications.
- Valencia, R.R. (2020). *International deficit thinking: Educational thought and practice*. Routledge.
- Weiner, L. (2003). Why is classroom management so vexing to urban teachers? *Theory into Practice*, 42(4), 305-312. [https://doi.org/10.1207/s15430421tip4204\\_7](https://doi.org/10.1207/s15430421tip4204_7)
- Yun, S.H., & Weaver, R.D. (2010). Development and validation of a short form of the attitude toward poverty scale. *Advances in Social Work*, 11(2), 174-187.
- Yurt, E. (2024). *Sosyal bilimlerde çok değişkenli analizler için pratik bilgiler - SPSS ve AMOS uygulamaları* [Practical information for multivariable analysis in social sciences - SPSS and AMOS applications]. Nobel Academy.
- Zammit, S. (2020). *Education for all learners: Elimination "deficit thinking" in favour of inclusive and culturally responsive schooling in Malta* [Doctoral dissertation, University of Lincoln]. ProQuest Dissertations & Theses Global.

## APPENDIX

## Appendix 1. Deficit Thinking Scale for teachers-Turkish version.

ÖĞRETMENLER İÇİN EKSİKLİK DÜŞÜNCESİ ÖLÇEĞİ						
Madde No	MADDELER Aşağıdaki ölçek formlarında yer alan cümleler için 1.Tamamen katılmıyorum, 2.Katılmıyorum, 3.Kararsızım, 4.Katılıyorum, 5.Tamamen katılıyorum şeklinde seçenekler sunulmuştur. Doğru veya yanlış cevabın bulunmadığı sadece derecelemenin yapıldığı formlarda görüşlerinize uygun olan cevabı (X) işareti ile işaretleyiniz. Lütfen maddeleri <b>BOŞ BIRAKMAYINIZ</b> .	Tamamen katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Tamamen katılıyorum
		<b>KURBANİ SUÇLAMA</b>				
<b>Sosyoekonomik düzeyi düşük olan öğrencilerin diğer öğrencilere göre;</b>						
1	Uygun olmayan arkadaş çevresine sahip olma ihtimali daha yüksektir.	1	2	3	4	5
2	Agresif davranışlar sergileme ihtimali daha yüksektir.	1	2	3	4	5
3	Disiplini bozan davranışlar sergileme ihtimali daha yüksektir.	1	2	3	4	5
<b>ÇEVRESİNİ SUÇLAMA</b>						
<b>Sosyoekonomik düzeyi düşük olan aileler diğer ailelere göre;</b>						
4	Veli toplantısına daha düşük katılım gösterir.	1	2	3	4	5
5	Çocuklarının akademik başarısına daha az ilgi gösterir.	1	2	3	4	5
6	Çocuklarının ders çalışmasını daha az teşvik eder.	1	2	3	4	5
7	Çocuklarının yaşadığı sorunlara karşı daha az duyarlıdır.	1	2	3	4	5
8	Çocuklarını ödevle ilgili konularda daha az destekler.	1	2	3	4	5
9	Çocuklarını toplumsal kuralları öğrenme konusunda daha az destekler.	1	2	3	4	5
10	Öğretmenlerle iletişim kurmakta daha fazla zorlanır.	1	2	3	4	5
<b>EĞİTİLEBİLİRLİK</b>						
<b>Sosyoekonomik düzeyi düşük olan öğrencinin diğer öğrencilere göre akademik başarısı;</b>						
11	Öğretmenin onunla birebir ilgilenmesinden daha fazla etkilenir.	1	2	3	4	5
12	Öğretmenin onu motive etmesinden daha fazla etkilenir.	1	2	3	4	5
13	Öğretmenin onunla iyi iletişim kurmasından daha fazla etkilenir.	1	2	3	4	5
14	Öğretmenin ona geri bildirim vermesinden daha fazla etkilenir.	1	2	3	4	5
15	Öğretmenin onunla ilgili akademik beklentilerinden daha fazla etkilenir.	1	2	3	4	5
<b>BASKI</b>						
<b>Eğitim sistemi sosyoekonomik düzeyi düşük olan öğrenciler için;</b>						
16	Gerekli öğrenme koşullarını sağlar.	1	2	3	4	5
17	Dezavantajları ortadan kaldıracak gerekli tedbirleri alır.	1	2	3	4	5
18	Başarılı olmalarını sağlayacak olanaklar sunar.	1	2	3	4	5
19	Gerekli öğrenme koşullarını sağlamakta yetersizdir.	1	2	3	4	5
<b>SAHTE BİLİM</b>						
<b>Sosyoekonomik düzeyi düşük olan öğrencilerin akademik başarısızlığı;</b>						
20	Öğrencinin gerekli çabayı göstermemesinden kaynaklanır.	1	2	3	4	5
21	Öğrencinin öğrenmeye istekli olmamasından kaynaklanır.	1	2	3	4	5
22	Öğrencinin eğitimi yeterince önemsememesinden kaynaklanır.	1	2	3	4	5

Note. Item 20 will be reverse-coded.

## The effect of STEM practices on students' attitudes and achievements: A meta-analysis study

Abdulkadir Kurt<sup>1\*</sup>, Muhammed Akıncı<sup>2</sup>

<sup>1</sup>Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Ağrı, Türkiye

<sup>2</sup>Recep Tayyip Erdoğan University, Faculty of Education, Department of Educational Sciences, Rize, Türkiye

### ARTICLE HISTORY

Received: Mar. 26, 2024

Accepted: July 27, 2025

### Keywords:

STEM education,  
Meta-analysis,  
Attitude,  
Achievement.

**Abstract:** This meta-analysis investigates the impact of STEM (Science, Technology, Engineering, and Mathematics) practices on students' academic achievement and attitudes toward STEM subjects. The study used meta-analysis as a research methodology and searched for relevant literature in databases such as EBSCOhost, Scopus, Web of Science, Google Scholar, and ULAKBIM. The keywords used were "STEM," "STEM Education," and "experimental studies on STEM." The search yielded 45 studies, including articles and proceedings papers, of which 22 met the inclusion criteria for the meta-analysis. The Comprehensive meta-analysis (CMA) program was used as a data analysis tool for the data obtained. Effect sizes were calculated, and the values obtained between variables in the studies were presented in the forest plots at a 95 percent confidence interval. As a result, the individual studies included in the analysis are heterogeneous, and the achievement level in the experimental group is approximately 5 points higher than in the control group. So, the achievement level of the students who received STEM education is (on average) 4.89 units higher than those who did not. The results showed that STEM education had a positive and significant impact on students' attitudes toward the course and academic achievement compared to other methods. Therefore, STEM education enhances students' attitudes towards the course and academic achievement.

## 1. INTRODUCTION

There are numerous areas where development and change can be discussed today. With each passing day, discoveries and innovations emerge that have the potential to revolutionize the way we live and work. The pace of advancement in science and technology is truly remarkable. In this context, scientific research enables the production of scientific knowledge to solve various problems and also contributes to fields such as technology, engineering, and mathematics. Regarding this matter, it is possible to delve into technological advancements by opening a separate parenthesis. In addition to the developments in information and communication technologies since the 2000s, research on nuclear fusion and artificial intelligence is making significant progress today. Significant developments in nuclear fusion studies are expected to provide a solution to the future energy crisis (BBC, 2022). Again, artificial intelligence platforms such as Dall-E and ChatGPT, developed by OpenAI, have been

\*CONTACT: Abdulkadir KURT ✉ [akurt@agri.edu.tr](mailto:akurt@agri.edu.tr) 📍 Ağrı İbrahim Çeçen University, Faculty of Education, Department of Educational Sciences, Ağrı, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

widely used quickly, and superior versions of these applications continue to be developed (OpenAI, 2023). These developments draw attention to the importance of science, technology, engineering, and mathematics combined with the acronym STEM. STEM education faces several obstacles, including the need for properly trained and qualified instructors, the need for students to be adequately prepared, the limited integration of STEM disciplines into educational programs, and more credible scientific research in this area. These are significant challenges that must be carefully considered to effectively address the needs of learners and contribute to the advancement of STEM fields (Kiazai *et al.*, 2019). For this reason, when promoting a comprehensive approach to STEM education, it is essential to prioritize the quality of both students and teachers. This case means considering not only cognitive aspects but also affective processes. It is crucial to conduct scientific research on this topic to understand how both types of processes impact the learning experience. At this point, attitude and success can be considered remarkable cognitive and affective elements regarding student quality. The objective of this paper was to systematise the existing literature on the relation between STEM practices and attitude and academic achievement of students. Instead of conducting a narrative literature review, a meta-analytic technique was applied. A standard literature review typically comprises a commentary on the findings of previous works. The character of this approach, however, is a qualitative one. It neither allows for a quantitative assessment of the effect of interest nor does it enable its standardisation for different methods applied across studies. The inability to compare estimates obtained by different researchers often leads to a substantial bias in the selection of the literature used in the review. These problems can be overcome by applying a meta-analysis, in other words, by conducting a quantitative literature review. As stated by Matysiak and Vignoli (2007), this methodology, relatively new in the social and educational sciences, has been developed to synthesise, combine, and interpret the abundance of empirical evidence on a specific topic. The purpose of this research is to analyse the effect of STEM practices on students' attitudes and academic achievement by combining and interpreting empirical evidence from related research topics.

### 1.1. STEM and Reflections on Education

STEM education represents a shift from traditional teaching methods toward a more interdisciplinary and applied approach. By combining scientific inquiry with artistic creativity, STEM aims to cultivate a broad skill set that includes critical thinking, collaboration, and innovation. Çepni and Çil (2009) state that individuals can gain high-level scientific process skills from school age. According to the data obtained in this study, it was determined that STEM practices caused improvement in students' scientific process skills. The fact that STEM practices have caused changes in scientific process skills demonstrates that high-level developments can be achieved through the development of teaching processes that integrate STEM. According to Yıldırım (2016), "STEM education is a contemporary approach that aims to use an integrative approach while educating individuals and teaching them the necessary skills in daily life and scientific process that can meet the needs by equipping them with 21st-century skills. STEM education is recognized as an integrative approach that enables individuals to connect their daily experiences with the course material. This education is crucial for individuals to stay informed about global developments, propose innovative solutions, and adapt to emerging trends. In this respect, STEM education has a structure that can be applied in each stage of education (Aydağül & Terzioğlu, 2014; Breiner *et al.*, 2012; Bybee, 2010). STEM education is essential for developing critical thinking, problem-solving skills, and preparing students for a future in an increasingly technology-driven world. Reflecting on STEM practices in education involves not only understanding its core components but also considering how these practices can be effectively implemented and evaluated to enhance learning outcomes (Belland *et al.*, 2017; Lestari *et al.*, 2018).



## 1.2. The Relationship Between STEM Education and Student Achievement

Research consistently shows that students involved in STEM programs perform better academically, especially in math and science. For instance, Özkan and Doğan (2022) found that seventh-grade students engaged in STEM activities demonstrated significantly higher academic performance and improved attitudes toward science. Scientific process skills are skills that can be developed through formal education processes in schools and can be influenced by the teaching methods and techniques employed in these processes. The integration of arts into STEM education nurtures creativity and problem-solving abilities. Henriksen (2014) emphasizes that excellent STEM teachers often incorporate creative strategies from the arts to enhance learning outcomes. Yigit and Bagci (2024) support this with a meta-analysis showing that STEM education significantly improves student creativity. STEM education utilizes hands-on, inquiry-based learning to enhance student engagement. Beers (2011) argues that 21st-century skills, such as communication and collaboration, are best developed through STEM learning environments. Yilmaz and Yilmaz (2024) found that gamified STEM activities increase both motivation and academic performance. At the same time, STEM education prepares students for real-world challenges by fostering critical thinking and collaboration. Marshall and Horton (2011) show that inquiry-based instruction, a core component of STEM, leads to higher-order thinking skills. Yakman (2008) presents a model for integrative education that aligns well with contemporary workforce demands.

## 1.3. The Relationship Between STEM Education and Student Attitude

Student attitude can be broadly defined as a learner's internal disposition toward education, including their emotional responses, cognitive beliefs, and behavioral intentions (Fraser, 1998). It is often reflected in students' enthusiasm for learning, willingness to engage in academic tasks, and their resilience in the face of challenges. Student attitude is a multifaceted construct encompassing learners' emotions, beliefs, and predispositions toward their academic environment and learning processes. As an integral component of student behavior, attitude has a significant influence on motivation, classroom engagement, and ultimately, academic success (Ajzen, 1991). Research consistently demonstrates a strong correlation between positive student attitudes and higher academic achievement (Pintrich & De Groot, 1990). Students with a growth mindset—those who believe that intelligence can be developed—tend to embrace challenges and persist in the face of setbacks (Dweck, 2006). Conversely, students with negative attitudes often experience lower motivation, increased absenteeism, and reduced performance. A constructive attitude not only enhances academic scores but also fosters critical thinking, creativity, and collaborative skills. Furthermore, in recent years, studies have shown that teaching processes with STEM practices can positively affect students' learning journeys, such as inner motivation and positive attitudes for classroom achievement (Bedar & Al-Shboul, 2020; Belland, *et al.*, 2017; Cunningham & Hester, 2007; Lestari, *et al.*, 2018; Mousoulides, 2013). Reflecting on STEM practices in education reveals that, while they offer transformative potential for student learning, they also present unique challenges. It requires teachers to be flexible, creative, and supportive in their approach. Additionally, STEM education should focus not only on knowledge acquisition, but also on developing lifelong skills that students can apply to any career or area of their life. By integrating hands-on learning, problem-solving, and interdisciplinary approaches, STEM education can better prepare students for the complexities of the modern world, fostering not just technical skills but also collaboration, critical thinking, and adaptability (Sanders, 2009; Xie *et al.*, 2015). STEM practices have a direct impact on student success and attitude by fostering a deeper, more engaging learning experience. Hands-on activities, inquiry-based learning, problem-solving, technology integration, and collaboration all play key roles in developing not only students' academic abilities but also their mindset and motivation. When STEM practices are implemented thoughtfully, they can help students succeed academically and develop positive attitudes toward learning, challenges, and their potential future in STEM fields (Madden *et al.*, 2016).

#### 1.4. The Significance and Challenges of STEM Education

Individuals may require some level of STEM education to understand the significance of scientific and technological advancements and their impact on human life, as well as to comprehend issues like global climate change, epidemics, environmental pollution, and water scarcity (Marrero, 2014). Furthermore, understanding STEM is crucial for individuals to make informed decisions that positively impact themselves, their families, and their communities (Tate *et al.*, 2012). In the context of contemporary society, the issue of sustainability is arguably one of the most pertinent. In this context, STEM education assumes a critical role in generating competent individuals who can proffer innovative solutions to this challenge. Ensuring that all students are equipped with an understanding and exposure to the fields of STEM is a crucial step towards fostering individual development and making a significant contribution to the global community. This effort can lead to an increase in the number of professionals in diverse fields such as engineering, medicine, science, and mathematics, which can positively impact the world at large (Blotnick *et al.*, 2018). Thus, integrating all members of society into STEM education is paramount in sharing diversified research and knowledge, ultimately leading to an enhanced innovation process fuelled by a broad range of perspectives and data (Marrero, 2014).

Initially, STEM education generated enthusiasm across a wide range of fields and became a topic of interest, from botany to industries producing consumer goods (Bybee, 2013). However, educators often encounter various challenges when teaching STEM subjects (Ejiwale, 2013; Martín-Páez *et al.*, 2019):

- Deficiencies in training qualified STEM teachers,
- Problems related to student readiness and motivation,
- Difficulty integrating fields such as technology and engineering into schools and curricula,
- Insufficient content for STEM education,
- Problems related to measurement and evaluation that are appropriate for STEM education,
- Laboratory and teaching environment problems for STEM education,
- Challenges to simplify the technical and complex issues concerning STEM,
- Limitations in research on STEM education,
- Difficulty converting the STEM concept from a slogan to an educational concept.

When examining competencies related to STEM education, researchers typically approach the subject in two different ways. Some studies examine students' attitudes and achievements towards STEM education in real-world settings (Beatty, 2011; Hackman, 2021; National Research Council, 2011; Vennix *et al.*, 2018). These studies primarily aim to provide descriptive insights. Other studies, on the other hand, employ experimental research methods to analyze more complex data and examine the impact of STEM education on student attitudes and achievement levels (Baran *et al.*, 2019; McClain, 2015; Tolliver, 2016; Wang *et al.*, 2022).

When executed efficiently, experimental research yields valid and reliable data on the variables being studied. However, due to its nature, this type of research is typically conducted on a small sample size in the educational field (Creswell, 2015). The issue at hand is the question of whether the results of these studies can be applied to the broader universe. To address this, researchers employ methods such as meta-synthesis or content analysis for qualitative data and meta-analysis for quantitative data to combine findings from similar studies and produce more universally applicable results (Cohen *et al.*, 2007). Since the related studies on STEM education are primarily experimental, meta-analysis can be considered a fundamental research type from which we can benefit.

The literature in this field includes relevant investigations. Some of these studies employed meta-synthesis and content analysis to examine existing research on STEM education (Kanadlı, 2019; Kaya, 2020; Ormancı, 2020; Yıldırım, 2016). A limited number of studies, utilizing meta-analysis as their approach, specifically focused on analyzing the different effects of certain

variables in experimental studies related to STEM education (Ayverdi & Aydın, 2020; Değer & Yapıcı, 2022; Kazu & Kurtoğlu Yalçın, 2021; Ulum, 2022). According to Olasehinde and Olatoye (2014), attitudes significantly influence academic performance across various subjects. However, no studies have been found that deal with variables such as attitude and achievement, which are essential cognitive and affective processes in STEM education, in an integrated way. In other words, various studies have assessed the effects of diverse integrated STEM studies. By compiling the results in the relevant literature, it is possible to draw a broad conclusion about the impact of different integrated STEM studies on student achievement and attitudes. From this perspective, it is considered essential to conduct a meta-analysis study that examines the impact of STEM education on students' attitudes and achievement levels in a comprehensive manner.

When examining the literature, it becomes apparent that STEM practices have an impact on student achievement and attitude in various areas. These can be summarized in [Table 1](#) below.

**Table 1.** *The effects of STEM practices on students' achievement and attitude in the literature.*

STEM and Student Achievement		STEM and Student Attitude	
Steps	Impacts	Steps	Impacts
Active Learning Approaches	It helps students develop critical thinking and problem-solving skills.	Fostering a Growth Mindset	This can foster a <b>growth mindset</b> , where students believe that their abilities can be developed through effort and perseverance.
Integration of Technology and Tools	This helps students understand complex scientific, mathematical, and engineering concepts in a more interactive way.	Promoting Curiosity and Intrinsic Motivation	It fosters intrinsic motivation, such as natural curiosity and interest in the subject matter.
Collaboration and Teamwork	It helps students learn from one another, refine their ideas, and develop teamwork skills.	Increased Self-Confidence	It can significantly boost students' <b>self-confidence</b> in their abilities, especially when they overcome challenges and see tangible results.
Interdisciplinary Approach	This approach helps students see the relevance of what they are learning	Equity and Inclusivity	It leads to greater diversity among future STEM professionals, regardless of their background.
		Real-World Relevance	It helps them contribute to solving global issues (e.g., climate change, technological innovation)

According to [Table 1](#), STEM practices offer a powerful means of fostering higher-order thinking, collaborative problem-solving, and technological fluency. When implemented thoughtfully, these methods can improve both student achievement and attitude, making learning more meaningful and engaging. This interdisciplinary approach supports the broader goal of preparing students for the complexities of modern life and work.

### 1.5. Aim of the Study

This research aims to conduct a meta-analysis on the effect of STEM education on students' attitudes and achievement levels. In the present research, individual studies that discussed the relationship between STEM education and attitudes and academic achievement of students in the industry will be examined via meta-analysis methodology. The hypothesis on the relationship between the variables and on the intensity of the relationship in question will be tested. For this purpose, the following research questions were included in the study:

- What are the achievement levels of students who attended STEM education according to

studies analysed?

- What are the attitudes of students who attended STEM education according to studies analysed?

## 2. METHOD

A meta-analysis was employed as the research methodology for this study. Meta-analysis is the process of systematically collecting, synthesizing, and analysing the findings of multiple studies on a specific subject (Shelby & Vaske, 2008). Meta-analysis is a method of combining the results of independent, multiple studies and performing statistical analysis of the obtained research findings (Parker *et al.*, 2013), explaining the results of each study with the help of a numerical index, and then combining those estimates throughout the studies to reach a summary (Quintana & Minami, 2006). By researching the sample selected through meta-analysis, researchers attempt to make predictions and generalizations about the population, acknowledging a certain probability of error. Meta-analysis provides a general effect size ( $r$ -value) and confidence interval on the cumulative evidence derived from the combination of two or more studies (Hedges & Pigott, 2004). The fixed-effect or randomized-effect model is used to analyse the studies in a meta-analysis. Suppose the results of individual studies in the meta-analysis are homogeneous. In that case, the fixed-effect model is preferred, but if the results are heterogeneous, then the random-effects model is selected to analyse the data. (Celiker *et al.*, 2019).

The methodological process discussed regarding the meta-analysis process is presented below under 11 subheadings as eligibility criteria, information sources, search strategy, selection process, data collection process, data items, study risk of bias assessment, effect measures, synthesis methods, reporting bias assessment, and certainty assessment (Page *et al.*, 2021).

### 2.1. Eligibility Criteria

Meta-analysis studies differ from most studies, and it is an analysis method that analyzes the results of individual studies conducted on the subject of interest. In the words of Glass (1976), the meta-analysis method, which is the synthesis of individual studies related to the researched topic and the presentation of an analysis from the beginning, aims to show the big picture. Therefore, in this study, the results of individual studies conducted on the researched topic are used. In addition to these, all of the individual studies previously conducted on the subject under consideration constitute the universe and therefore the sample of the meta-analysis (Tarım, 2003). Those meeting the inclusion criteria below were included in the meta-analysis:

- Full-text articles that examine the relationship between student achievement/attitude and STEM education.
- Publications subjected to peer review and have been published in academic journals.
- Studies with correlation coefficients to get the standardized effect size in the meta-analysis studies.

### 2.2. Information Sources

A literature search was made in EBSCOhost, Scopus, Web of Science, Google Scholar, and ULAKBİM (Turkish Academic Network and Information Center) databases using the keywords “STEM,” “STEM Education,” and “experimental studies on STEM.” The databases preferred in the study are widely accepted indexes, especially in the field of educational sciences. However, since academic publications that passed the review process were preferred, theses were not included in the study. For the selection of studies, criterion sampling was employed in this study, rather than the traditional purposive sampling. In purposive sampling, researchers select samples that they believe will meet their information needs, depending on the study's purpose (Fraenkel & Wallen, 2009). Criterion sampling is the method of using people, events, or objects with certain predetermined qualities in a research sample selection

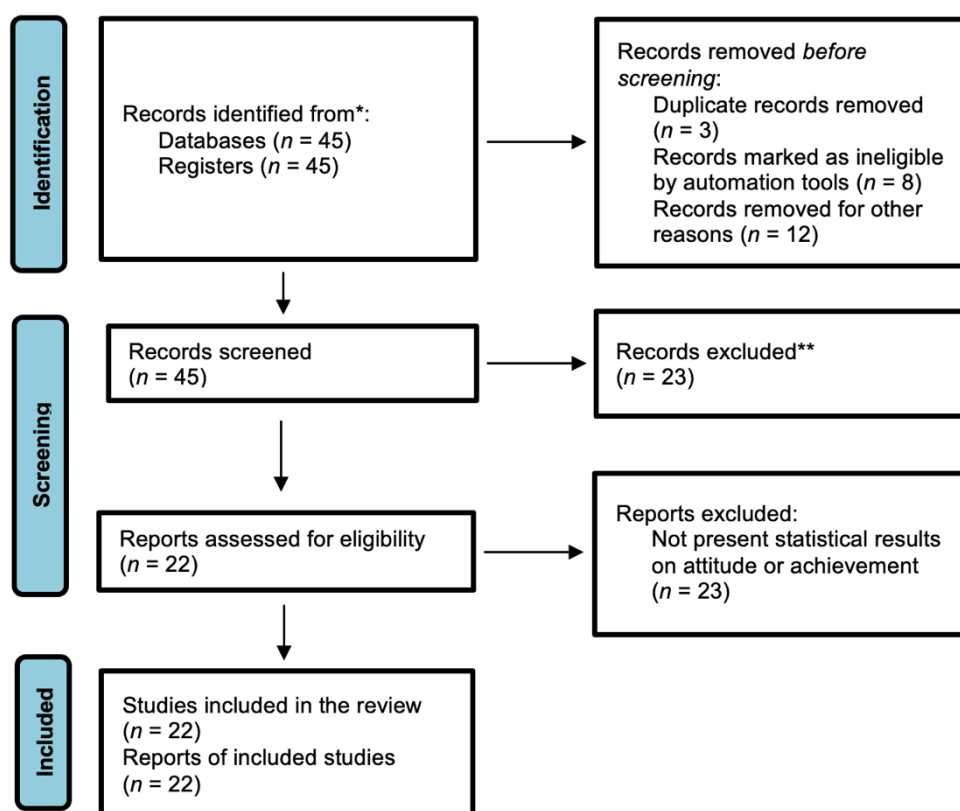
(Büyüköztürk *et al.*, 2015). At the end of the literature search, 45 studies were collected, comprising articles and proceedings papers.

### 2.3. Search Strategy

The 22 studies included in the analysis (Appendix A) generally examined the effect of STEM education on student achievement and attitude. These studies were included in the analysis process as they provided experimental data as a criterion.

### 2.4. Selection Process

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart is a key component of systematic review methodology, providing a transparent and structured way to report the process of study selection. The flowchart outlines the different stages of the study selection process, as outlined by Moher *et al.* (2009). The study screening process of the research is presented in a flowchart of the meta-analysis process (see, Figure 1).



**Figure 1.** PRISMA flowchart of the study screening process (Adapted from Moher *et al.* (2009).

As seen in the flowchart:

While 45 records were initially identified through database searches. 23 of these records were excluded based on the title/abstract screening. Then, 22 full-text articles were assessed for eligibility, and 23 were excluded. Ultimately, 22 studies were included in the meta-analysis.

### 2.5. Data Collection Process

Two researchers coded the titles, author(s) of the study, publication years, publication types, sampling size, correlation coefficients, populations, and scales. After the results were obtained, the Kappa statistic was used for inter-coder reliability. Kappa statistic, which is frequently used to determine inter-rater reliability, was developed to determine the degree of agreement between two raters scoring at the classification level. Kappa statistics take values between -1 and +1. It is stated that the closer the value ranges for the interpretation of Kappa Statistics are to +1, the higher the inter-coder reliability. The result obtained according to the inter-coder



scoring was determined as .91, and this was interpreted as the inter-coder reliability index is very high (Landis & Koch, 1977)

## 2.6. Data Items

In meta-analysis, data item selection refers to the process of choosing specific pieces of data from individual studies that will be combined and analysed to assess the overall effect of an intervention. Careful selection of data items is crucial to ensure that the meta-analysis yields reliable and valid conclusions. Therefore, the research questions determine which data items are relevant before selecting them. From the eligible studies, researchers include specific data items for meta-analysis. The key data items include: Effect sizes (e.g., mean difference, odds ratio, risk ratio, hazard ratio, standardized mean difference).

- Sample sizes for experimental and control groups.
- Outcome measures (e.g., means, standard deviations, confidence intervals).
- *p*-values or other statistical values indicating significance.

## 2.7. Study Risk of Bias Assessment

Risk of bias assessment is an essential component of meta-analysis because it helps determine the reliability and validity of the studies included. Bias refers to any systematic error that can distort the true effect of an intervention or treatment. Identifying and evaluating the risk of bias within the individual studies is crucial for drawing accurate conclusions from the meta-analysis. In this study, researchers used statistical techniques to adjust for bias in meta-analysis, such as funnel plot analysis to check for publication bias. Funnel plot, Classic fail-safe *N*, Begg and Mazumdar Rank Correlation, Egger regression, and Duval Tweedie's trim-and-fill methods were used to determine whether the studies included in the meta-analysis caused any publication bias.

## 2.8. Effect Measures

In meta-analysis, effect measures (or effect sizes) are used to quantify the magnitude of the effect of an intervention or treatment across different studies. These measures summarize the relationship between an intervention and an outcome, making it possible to combine results from different studies in a meaningful way. The random-effects model is used to combine the results of different studies, accounting for the fact that the true effect size might differ from study to study due to differences in study populations, interventions, and methodologies. This model assumes that the effect sizes estimated in each study are not identical but rather vary around a central true effect. Therefore, the data must be converted into a standard unit of measurement to statistically combine the individual research findings and reach a consensus in meta-analysis studies. The effect size index in a correlation study is calculated as the correlation between the independent variable classification and the individual scores of the dependent variable (Neely *et al.*, 2010). The effect sizes obtained from the test statistics of individual studies were standardized and tested to determine the strength of the relations in the context of the specified hypothesis in this study. Fisher's *Z* formula was used to calculate the effect size; also, correlation-based effect size classification was used to interpret the effect size obtained (Cohen, 2007):

- Effect size < 0.10 : very low level
- $0.10 \leq \text{effect size} < 0.30$  : low level
- $0.30 \leq \text{effect size} < 0.50$  : medium level
- $0.50 \leq \text{effect size} < 0.80$  : strong level
- Effect size  $\geq 0.80$  : very strong level

## 2.9. Synthesis Methods

In meta-analysis, synthesis methods refer to the statistical techniques used to combine results from multiple studies and draw a comprehensive conclusion. These methods allow researchers

to integrate findings from diverse studies, quantify the overall effect size, and account for study differences or heterogeneity. Statistical heterogeneity is related to the variability in effect sizes in individual studies. It is known that only if there is real heterogeneity between the estimated effect sizes of several studies, as meta-analyzed, is it clearly visible (Huedo-Medina *et al.*, 2006). In meta-analysis, heterogeneity exists when the variance between individual studies is significantly increased. Heterogeneity tests and heterogeneity measures are not directly related to the variance value between individual studies, but rather to the increased variance value due to heterogeneity (Mittlböck & Heinzl, 2006; Sutton *et al.*, 2000). In this research, Cochran's Q statistic is used for the heterogeneity test. It is the most common and straightforward approach used to assess whether there is real heterogeneity among the individual studies included in the meta-analysis (Cochran, 1954).

## 2.10. Reporting Bias Assessment

There are several methods used to assess the potential for reporting bias in a meta-analysis. The main strategies involve visual inspection of data and statistical tests to determine whether smaller studies or studies with certain characteristics (e.g., larger effects) are more likely to be published. In this research, a funnel plot was used to assess publication bias by plotting the effect size estimates from individual studies. Egger's test was used to formally assess funnel plot asymmetry. This test was used because it can help confirm whether the asymmetry is statistically significant.

## 2.11. Certainty Assessment

Certainty assessment (also referred to as "quality of evidence" or "confidence in estimates") in meta-analysis is a process that evaluates the degree of confidence one can have in the overall findings of the meta-analysis. This is crucial because, while meta-analysis aggregates the results of multiple studies, the strength of the conclusions depends on factors such as the study designs, consistency of results, risk of bias, and other methodological considerations. Six steps were followed in this research to assess certainty.

Step 1 (Evaluate risk of bias): This step evaluates the extent to which the individual studies included in the meta-analysis were well-designed and free from systematic errors

Step 2 (Assess heterogeneity): Heterogeneity is typically assessed in this step. Inconsistency measures how much the study results vary across different studies. If studies report conflicting findings or the results are highly variable, the certainty of the overall evidence is reduced.

Step 3 (Check for Indirectness): Indirectness refers to whether the evidence is directly applicable to the question being asked in the meta-analysis. It involves the degree to which the population, intervention, comparator, and outcomes in the studies match the research question.

Step 4 (Assess Imprecision): Imprecision refers to the extent to which the estimates of effect sizes are precise. This is typically evaluated by looking at the confidence intervals of the pooled effect estimate.

Step 5 (Publication bias): Publication bias refers to the tendency for studies with positive or significant findings to be more likely to be published than those with null or negative results. This can skew the results of a meta-analysis.

Step 6 (Final judgement): After considering the above factors, a final certainty rating is assigned to the overall body of evidence. The grade system typically rates the evidence as high, moderate, low, or very low certainty

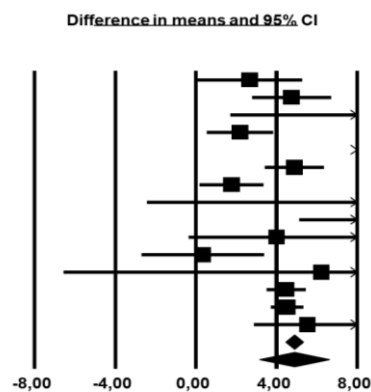
## 3. RESULTS

### 3.1. The Effect of STEM Education on Student Achievement

The effect of STEM Education on student achievement is given in [Table 2](#) and [Figure 2](#).

**Table 2.** Model statistics for each study.

Study	Difference in means	Standard error	Variance	Lower limit	Upper limit	z-value	p-value
Study (1)	2.68	1.30	1.70	0.12	5.24	2.05	.04
Study (2)	4.75	0.98	0.96	2.83	6.67	4.85	.00
Study (3)	8.04	3.22	10.38	1.73	14.37	2.50	.01
Study (4)	2.19	0.82	0.66	0.57	3.81	2.66	.01
Study (5)	12.00	0.68	0.47	10.67	13.33	17.63	.00
Study (6)	4.89	0.73	0.54	3.45	6.33	6.68	.00
Study (7)	1.77	0.79	0.62	0.22	3.32	2.24	.03
Study (8)	10.35	6.51	42.40	-2.41	23.11	1.59	.11
Study (9)	15.27	5.15	26.57	5.17	25.37	2.96	.00
Study (10)	4.01	2.20	4.85	-0.30	8.32	1.82	.07
Study (11)	0.34	1.53	2.35	-2.66	3.34	0.22	.82
Study (12)	6.22	2.61	6.84	1.10	11.34	2.37	.02
Study (13)	4.48	0.48	0.23	3.55	5.41	9.40	.00
Study (14)	3.50	0.66	0.43	2.22	4.78	5.31	.00
Study (15)	5.54	1.34	1.81	2.91	8.18	4.12	.00
Fixed	4.89	0.22	0.05	4.46	5.32	22.26	.00
Random	4.89	0.88	0.77	3.17	6.61	5.57	.00

**Figure 2.** Forest plot of experimental and control groups.

As seen in Table 2 and Figure 2, the combined effect according to the fixed and random effects model is 4.89. In other words, the achievement level in the experimental group is approximately 5 points higher than in the control group. In other words, the achievement level of the students who received STEM education is (on average) 4.89 units higher than those who did not. According to the data in Table 3, the individual studies included in the analysis are heterogeneous  $Q(154.880, p < .05)$  and  $I^2(90.96)$ , so the random effects model will be preferred in the meta-analysis.

**Table 3.** Variation in effect size.

Heterogeneity			Tau-squared			
df(Q)	p-value	I-squared	Tau Squared	Standard Error	Variance	Tau
14	.000	90.961	8.188	5.032	25.317	2.862

According to the random effects model in Table 4, the summary effect has a  $z$  value of 5.57 and a  $p$  value of .000 ( $p < .05$ ). Thus, the null hypothesis claiming that there was no real mean difference between the experimental and control groups was rejected, and it was concluded that the achievement levels of the students who received STEM education created a statistically significant difference compared to those who did not.

**Table 4.** Average effect size.

Model	Number of Studies	Effect size and 95% confidence interval					Test of null (2-Tail)	
		Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	z-value	p-value
Fixed	15	4.889	0.220	0.048	4.459	5.319	22.261	.000
Random	15	4.890	0.877	0.770	3.171	6.610	5.574	.000

### 3.2. Publication Bias Tests

Funnel plot, Classic fail-safe  $N$ , Begg and Mazumdar Rank Correlation, Egger regression, and Duval Tweedie's trim-and-fill publication bias tests were performed to test whether the individual studies included in the meta-analysis carried publication bias.

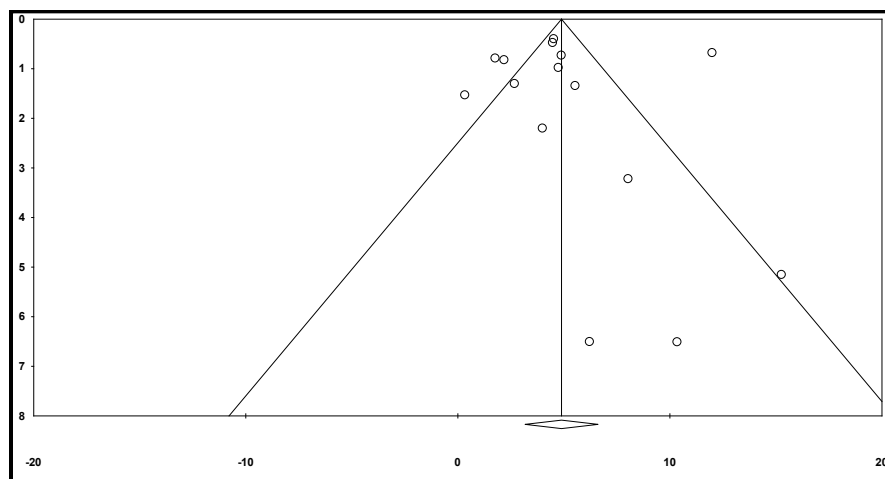
**Figure 3.** Funnel scatter plot.

Figure 3 shows the funnel plot test results of the individual studies included in the meta-analysis. In order to avoid publication bias, the circles representing each study are expected to be symmetrical to each other and gather at the top of the graph. An essential advantage of this method is that it provides the best unbiased effect size value. When Figure 3 is analysed, it is observed that 2 of the individual studies are located at the bottom of the graph, which may lead to publication bias. In order to make a general inference for publication bias, other publication bias tests need to be performed (see, Table 5).

**Table 5.** Publication bias tests.

Classic Fail-Safe $N$	Egger regression ( $p$ -value 2-tailed)	Kendall $Tau\ b$	Duval Tweedie's trim and fill (random effect)	
			Studies trimmed (to the right)	SMD observed (adjusted)
1297	.92	0.28	0	4.88-4.88

Classic fail-safe  $N$  refers to the number of new studies required to convert the overall probability value from the pooled test to a value greater than the specified critical value for statistical significance. That is, Classic fail-safe  $N$  calculates the number of missing studies, i.e., studies excluded in a meta-analysis. According to the Classic fail-safe  $N$  test results, the number of studies required for the  $p$ -value to be greater than .05 is calculated. Accordingly, it can be stated that to address publication bias in this meta-analysis study ( $p < .05$ ), 1297 more studies should be added to the analysis unit. Since it is not possible to reach this number of studies in this research area, this result is proof that there is no publication bias.

The Egger regression test, which determines the asymmetry in the funnel plot, indicates that there is no publication bias when the  $p$ -value is above .05. According to the Egger regression

test results, the  $p$ -value was above .05 (.92), indicating that our meta-analysis study does not carry publication bias.

A formal test for publication bias can be performed by examining the correlation between the effect estimates and their variances. The Begg and Mazumdar rank correlation test is a popular technique for assessing the likelihood of publication bias, which complements the funnel plot. In this method, the Kendall tau b coefficient is calculated. Without publication bias, this coefficient is expected to be close to 1, and the two-tailed  $p$ -value is not expected to make a significant difference. When the statistical values obtained as a result of the bias test are analysed ( $Tau\ b = 0.28$ ;  $p$ -value (two-tailed) .13,  $p > .05$ ), the evidence that there is no publication bias in the study is supported.

Duval Tweedie's trim and fill method is also used to estimate the possible number of missing studies in the meta-analysis and their impact on the overall findings. According to the adjusted  $SMD$  results of the truncated studies, no differences in the size and direction of the variables that could lead to publication bias were detected.

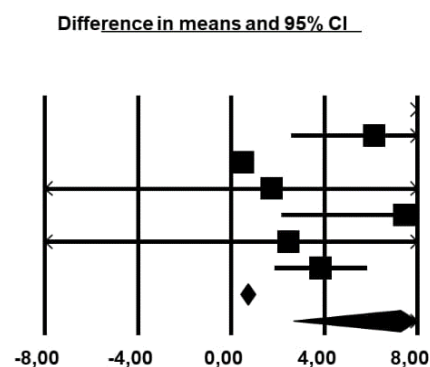
When the evidence obtained from the publication bias tests is evaluated in general, it can be stated that the meta-analysis study does not have an analysis unit that may lead to publication bias; in other words, the study unit included in the analysis does not carry publication bias.

### 3.3. The Effect of STEM Education on Student Attitudes

The effect of STEM Education on student attitudes is given in Table 6. As seen in the table and Figure 4, the combined effect is 0.72 according to the fixed-effects model and 7.28 according to the random-effects model.

**Table 6.** Model statistics for each study.

Study	Difference in means	Standard error	Variance	Lower limit	Upper limit	z-value	p-value
Study (1)	28.00	3.37	11.38	21.39	34.61	8.30	.00
Study (2)	6.13	1.81	3.27	2.59	9.67	3.39	.00
Study (3)	0.51	0.16	0.03	0.19	0.83	3.16	.00
Study (4)	1.73	7.08	50.06	-12.14	15.60	0.24	.81
Study (5)	7.44	2.69	7.23	2.17	12.71	2.77	.01
Study (6)	2.46	6.96	48.44	-11.18	16.10	0.35	.72
Study (7)	3.84	1.00	1.00	1.88	5.80	3.85	.00
Fixed	0.72	0.16	0.03	0.41	1.03	4.57	.00
Random	7.28	2.35	5.52	2.67	11.88	3.10	.00



**Figure 4.** Forest plot of experimental and control groups.

According to the data in Table 7, the individual studies included in the analysis are heterogeneous. According to  $Q$  (92.109,  $p < .05$ ) and  $I^2$  (93.48) values, the random effects model will be preferred in meta-analysis.



**Table 7.** Variation in effect size.

Heterogeneity			Tau-squared			
$df(Q)$	$p$ -value	I-squared	Tau Squared	Standard Error	Variance	Tau
6	.000	93.486	18	30.943	957.487	5303

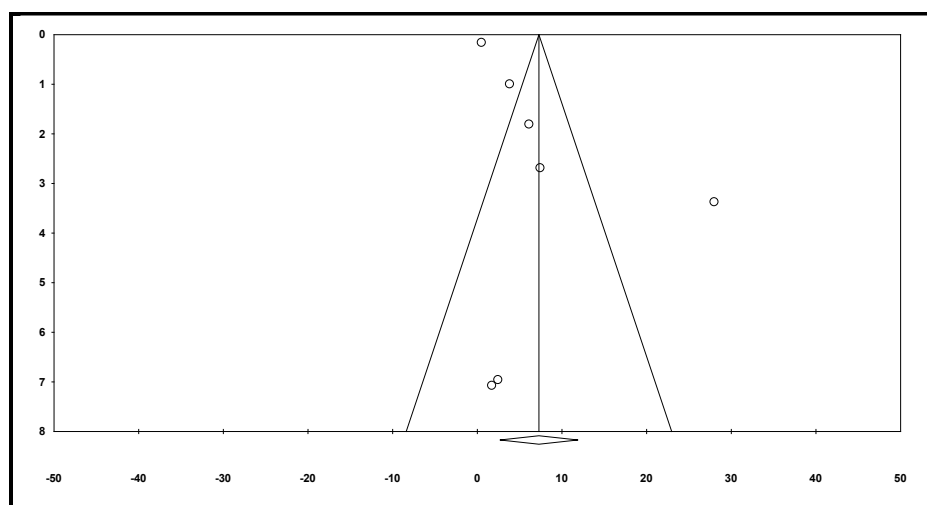
According to the random effects model in Table 8, the summary effect has a  $z$  value of 3.10 and a  $p$ -value of .00 ( $p < .05$ ). Thus, the null hypothesis claiming that there was no real mean difference between the experimental and control groups was rejected, and it was concluded that the attitudes of the students who received STEM training were statistically significantly different from those of the students who did not receive STEM training.

**Table 8.** Average effect size.

Model	Number of Studies	Effect size and 95% confidence interval					Test of null (2-Tail)	
		Point Estimate	Standard Error	Variance	Lower Limit	Upper Limit	$z$ -value	$p$ -value
Fixed	7	0.723	0.158	0.025	4.413	1.033	4.567	.000
Random	7	7.279	2.349	5.519	2.674	11.883	3.098	.002

### 3.4. Publication Bias Tests

Figure 5 shows the funnel plot test results of the individual studies included in the meta-analysis. When Figure 2 is analysed, it is observed that individual studies are not symmetrical in a way that may lead to publication bias and are located at the bottom of the graph. In order to make a general inference for publication bias, other publication bias tests need to be performed (see, Table 9).

**Figure 5.** Funnel scatter plot.**Table 9.** Publication bias tests.

Classic Fail-Safe $N$	Egger regression ( $p$ -value 2-tailed)	Kendall $Tau b$	Duval Tweedie's trim and fill (random effect)	
			Studies trimmed (to the right)	SMD observed (adjusted)
120	.06	-0.09	0	0.72-0.72

Based on the results of the Classic fail-safe  $N$  test, it can be inferred that an additional 120 studies would be required in the analysis to indicate the presence of publication bias in this meta-analysis ( $.05 < p$ ). Since it is not possible to reach this number of studies in this research area, this result shows that there is no publication bias. According to the Egger regression test

results, the  $p$ -value was found to be above .05 (.06), and according to this result, our meta-analysis study does not carry publication bias.

The Begg and Mazumdar rank correlation test shows that the Kendall tau  $b$  coefficient takes a negative value quite far from 1. However, since the  $p$ -value was not statistically significant ( $Tau\ b = -0.09$ ;  $p$ -value (two-tailed) .72,  $p > .05$ ), it can be interpreted that there is no publication bias in the study. When the results obtained from Duval Tweedie's trim and fill method are analyzed, the adjusted  $SMD$  results of the truncated studies reveal that there is no difference in the dimensions and directions of the variables that may lead to publication bias.

When the evidence obtained from the publication bias tests is evaluated in general, it can be stated that all tests except the funnel plot test show that the meta-analysis study does not have an analysis unit that may lead to publication bias; in other words, the study unit included in the analysis does not carry publication bias.

#### 4. DISCUSSION and CONCLUSION

The meta-analysis method has emerged as a comprehensive literature review involving the use of a systematic approach and statistical formulae. It is an undeniable fact that each of the scientific studies, which are the products of the research carried out by individual researchers to contribute to science with intensive labour, is very valuable (Aksoy Kürü, 2021). This meta-analysis study aims to examine the effect of STEM education on students' attitudes and achievement levels. Based on the calculations, it has been concluded that in the 22 studies included in the meta-analysis (15 related to academic achievement; 7 related to student attitudes), the effect of STEM education on students' attitudes and achievement towards the course is more positive compared to other methods according to the fixed effects model. In addition, since the effect size value was greater than 0.80, it has been determined to have a high effect level compared to Cohen's classification (Cohen, 2007). According to the results obtained, STEM practices have a positive impact on students' overall achievement levels. It was revealed that there was an individual difference in the scores. Since the learning of individuals depends on the differentiation in their minds, it is considered normal that the amount of increase in scores is at different levels. Nadelson *et al.* (2015) argue that the most important determinant of success in design courses is the individual learning capacity of the student. Similar results were found in the studies on STEM applications in the literature. For example, Abdelrahman and Asan (2006) stated that the design processes that students continue with their own experiences increase academic achievement, and Barker *et al.* (2010) stated that students' content knowledge changed positively in favour of the post-test in STEM applications. Wendell and Rogers (2013) stated that according to the results of STEM research, a significant increase was observed in the achievement level of students compared to the control group. In their study, Yıldırım and Altun (2015) also stated that academic achievement increased in courses conducted with a STEM approach. In addition to these results, Navruz *et al.* (2014) stated in their study that success in the teaching processes that students encounter newly cannot always be at a high level and positive. This finding suggests that STEM education has a highly positive impact on students' attitudes towards the related course and academic achievement compared to other methods. This finding suggests that STEM education enhances students' attitudes towards the course and academic achievement. In STEM education practices, students learn theoretical information in a more engaging and interactive manner, rather than in a monotonous way, through the STEM education model. STEM increases the creativity of the individual by developing innovative thinking. When mathematics, engineering and technology fields are combined by placing the science course in the focus, a permanent knowledge pool is formed in the lives of students. Similarly, Yamak *et al.* (2014) reported that STEM education improved students' scientific process skills, such as identifying and defining problems, researching, questioning, and solving problems, and that the designs created by students in the classroom helped them develop a positive attitude towards science classes. Additionally, Strong (2013)

observed that the engineering design process applied to elementary school students improved their scientific process skills.

According to Cohen (2007), the effect size classifications indicate that the impact of STEM education on students' attitudes towards the course and academic achievement levels has been determined to have a high and positively oriented effect size. This finding indicates that STEM education has a positive effect on students' attitudes towards the course and academic success level. Upon reviewing the literature, it is evident that STEM education has a positive impact on students' attitudes towards the course and academic achievement (Abanoz, 2020; Atik, 2019; Aydın, 2019; Bal, 2018; Borenstein *et al.*, 2009; Breiner *et al.*, 2012; Kalyoncu, 2021; Tatli, 2022). On the other hand, the meta-analysis study conducted by Değerli (2021) concluded that STEM education has a positive and broad impact on developing students' scientific process skills. This finding aligns with the research results. The study conducted by Lestari *et al.* (2018) indicated that STEM activities have a positive impact on students' scientific process skills. The meta-analysis conducted by Kim *et al.* (2018) also revealed that STEM education contributes to students' higher-level skills. In fact, research conducted by Bircan and Çalışıcı (2022) determined that STEM applications not only positively affect scientific process skills but also have an impact on 21st-century skills, including critical thinking, collaboration, communication, creativity, and sharing. In the meta-synthesis study on STEM education conducted by Herdem and Ünal (2018), positive effects on students' scientific process skills, academic achievement, attitudes, career awareness, and engineering processes were highlighted. The current scenario further corroborates the findings of the research. According to the findings obtained from the meta-analysis research conducted by Kazu and Kaplan (2024), it was determined that the effect of STEM applications on science process skills was positive and at a moderate level ( $g = 0.992$ ). This circumstance currently reinforces STEM education's positive and beneficial impact on students' attitudes towards the course and academic achievement levels. Similarly, Yamak *et al.* (2014) concluded that STEM education leads to improvements in students' scientific process skills and helps them develop positive attitudes towards science. Saçan (2018) found that the STEM-based curriculum improves the scientific process skills of seventh-grade students, increases their motivation towards STEM, and positively affects their attitudes towards socio-scientific issues. Toma and Greca (2018) observed that the integrative STEM learning model based on inquiry helped students to develop positive attitudes towards science and increased their academic success.

Apart from student achievement and attitude, the increasing importance of STEM in the global economy and society has brought these disciplines to the forefront of educational reform. Policymakers and educators are focusing on how best to prepare students for a future in a world shaped by technology, innovation, and complex global challenges. In this context, it is essential to consider both policy and practice implications for improving STEM education. So, both policy and practice play critical roles in shaping the future of STEM education. Policymakers need to prioritize funding, curriculum reforms, teacher preparation, and inclusivity to ensure that all students are equipped with the necessary skills for success in STEM fields. At the same time, educators must implement innovative and interdisciplinary teaching strategies, integrate technology, and provide opportunities for students to develop both technical and soft skills. The synergy between policy and practice will ultimately foster a robust and inclusive STEM education system that prepares students for the challenges and opportunities of the future.

The analysis reveals that STEM practices have not only a generally positive effect on students' academic achievement, motivation, problem-solving skills, and engagement but also provide critical evidence to inform educational practice and policy development. These insights are essential for ensuring that STEM education reforms are not only innovative but also grounded in empirical data that reflect local needs and realities. Based on these findings, several instructional recommendations are presented, particularly regarding curriculum development, teacher training, classroom practices, assessment methods, and policy implications. Findings

from the meta-analysis suggest that integrated STEM education significantly enhances students' academic outcomes (Çetin & Türkan, 2020). Therefore, curriculum designers should promote cross-disciplinary approaches that merge scientific inquiry with technological and mathematical applications. Interdisciplinary units and thematic instruction can create more coherent and engaging learning experiences (Kelley & Knowles, 2016). On the other side, the positive effects of STEM programs are especially prominent when learning is experiential and student-centered. Therefore, project-based, inquiry-based, and problem-based learning approaches should be prioritized. These strategies align with constructivist principles and foster deeper understanding through hands-on exploration (Capraro & Slough, 2013). The meta-analytic findings offer a robust foundation for improving both practice and policy in STEM education in Turkey. When translated into actionable strategies, these results can support a more effective, equitable, and future-ready education system aligned with national development goals.

As a result, the effects of STEM applications on students' academic achievement and attitudes were examined, and different effects other than these were excluded from the scope of the study. Researchers planning to conduct a study are advised to examine current studies that apply STEM education to various courses and materials, utilizing technology to support STEM education. Within the scope of STEM education, meta-analysis studies can be carried out on different subjects, such as their effects on various factors, including motivation and retention. The possible causes of regional differences in statistics attitudes-achievement relationships should be explored. Meta-analyses that include all statistics attitudes-achievement research, regardless of the attitude survey used, should be conducted. Finally, recognizing the limitations of the study is crucial for interpreting the findings with appropriate caution. Nevertheless, the current meta-analysis provides a meaningful synthesis of the available evidence on STEM education in Turkey and offers practical insights for educators and policymakers. So, future research should aim for more standardized reporting, include longitudinal data, and explore the role of contextual factors in greater depth.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

### Contribution of Authors

**Abdulkadir Kurt:** Resources, Methodology Research design, Visualization, Data collection, Data Analysis Writing-original draft. **Muhammed Akıncı:** Resources, Methodology Research design, Data collection, Supervision, Critical Review.

### Orcid

Abdulkadir Kurt  <https://orcid.org/0000-0002-4557-1179>

Muhammed Akıncı  <https://orcid.org/0000-0002-5001-2080>

### REFERENCES

- Abanoz, T. (2020). *STEM yaklaşımına uygun fen etkinliklerinin okul öncesi dönem çocuklarının bilimsel süreç becerilerine etkisinin incelenmesi* [The effect of STEM-based science activities on preschool children's scientific process skills] [Unpublished doctoral dissertation]. Gazi University.
- Abdelraheem, A., & Asan, A. (2006). The effectiveness of inquiry-based technology-enhanced collaborative learning environment. *International Journal of Technology in Teaching and Learning*, 2(2), 65–87.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211.

- Aksoy Kürü, S. (2021). *Meta-analiz* [Meta-analysis]. *Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 42, 215–229.
- Altunel, M. (2018). *STEM eğitimi ve Türkiye: fırsatlar ve riskler* [STEM education and Türkiye: Opportunities and risks]. *Seta Perspektif*, 207, 1–7.
- Atik, A. (2019). *STEM etkinliklerinin bilimsel süreç becerileri üzerine etkisi: 5 yaş örneği* [The effect of STEM activities on scientific process skills: The case of 5-year-old children] [Unpublished master's thesis]. Trabzon University.
- Ayaz, M.F., & Sekerci, H. (2015). The effects of the constructivist learning approach on students' academic achievement: a meta-analysis study. *Turkish Online Journal of Educational Technology*, 14 (4), 143-156.
- Aydağül, B., & Terzioğlu, T. (2014). Importance of science, technology, engineering and mathematics. *TÜSİAD Journal of Opinion*, 85, 13-19.
- Aydın, T. (2019). *STEM uygulamalarının okul öncesi öğrencilerinin bilimsel süreç becerileri ve bilişsel alan gelişimlerine etkisi* [The effect of STEM applications on preschool students' scientific process skills and cognitive domain development] [Unpublished master's thesis]. Fırat University.
- Ayverdi, L., & Aydın, S. Ö. (2020). *STEM eğitiminin akademik başarı üzerindeki etkisini inceleyen çalışmaların meta-analizi* [Meta-analysis of studies examining the effect of STEM education on academic success]. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, 14(2), 840–888.
- Bal, E. (2018). *FeTeMM (Fen, teknoloji, mühendislik, matematik) etkinliklerinin 48–72 aylık okul öncesi çocuklarının bilimsel süreç ve problem çözme becerileri üzerindeki etkisinin incelenmesi* [The effect of STEM activities on the scientific process and problem-solving skills of preschool children aged 48–72 months] [Unpublished master's thesis]. Marmara University.
- BBC. (2022, December 13). *Breakthrough in nuclear fusion energy announced*. <https://www.bbc.com/news/science-environment-63950962>
- Beatty, A.S. (2011). *Successful STEM education*. National Academies Press.
- Bedar, R.W. A.-H., & Al-Shboul, M.A. (2020). The effect of using STEAM approach on motivation towards learning among high school students in Jordan. *International Education Studies*, 13(9), 48–57.
- Beers, S.Z. (2011). 21st century skills: Preparing students for their future. *Kappa Delta Pi Record*, 47(2), 64-67. <https://doi.org/10.1080/00228958.2011.10516575>
- Belland, B.R., Walker, A.E., Kim, N.J., & Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: A meta-analysis. *Review of Educational Research*, 87(2), 309-344.
- Bircan, M.A., & Çalışıcı, H. (2022). *STEM eğitimi etkinliklerinin ilköğretim dördüncü sınıf öğrencilerinin STEM'e yönelik tutumlarına, 21. yüzyıl becerilerine ve matematik başarılarına etkisi* [The effects of STEM education activities on fourth-grade students' attitudes towards STEM, 21st-century skills, and mathematics achievement]. *Eğitim ve Bilim Dergisi*, 47(211), 87–119. <https://doi.org/10.15390/EB.2022.10710>
- Blotnick, K.A., Franz-Odenaal, T., French, F., & Joy, P. (2018). A study of the correlation between STEM career knowledge, mathematics self-efficacy, career interests, and career activities on the likelihood of pursuing a STEM career among middle school students. *IJ STEM*, 5(22).
- Borenstein, M., Hedges, L.V., Higgins, J.P.T. and Rothstein, H.R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, West Sussex.
- Breiner, J.M., Harkness, S.S., Johnson, C.C., & Koehler, C.M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School science and mathematics*, 112(1), 3–11.
- Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, E., Karadeniz, Ş., & Demirel, F. (2015). *Bilimsel araştırma yöntemleri* [Scientific research methods]. Pegem Akademi Yayıncılık.



- Bybee, R.W. (2013). *The case for STEM education: Challenges and opportunities*. NSTA Press.
- Capraro, R.M., & Slough, S.W. (Eds.). (2013). *Project-based learning: An integrated science, technology, engineering, and mathematics (STEM) approach*. Springer.
- Celiker, N., Ustunel, M.F. and Guzeller, C.O. (2019). The relationship between emotional labour and burnout: A meta-analysis. *Anatolia*, 30(3), 328-345. <https://doi.org/10.1080/13032917.2019.1581625>
- Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1), 101-129.
- Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. Routledge.
- Cohen, L. (2007). Experiments, quasi-experiments, single-case research and Meta-analysis. In L. Cohen, L. Manion, & K. Morrison, (Eds.), *Research Methods in Education* (pp.272–296). Routledge.
- Creswell, J.W. (2015). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (5th Edition). Pearson Prentice Hall.
- Cunningham, C.M., & Hester, K. (2007, March). Engineering is elementary: An engineering and technology curriculum for children [Oral Proceeding]. *American Society for Engineering Education Annual Conference & Exposition*, Honolulu, HI.
- Çepni, S., & Çil, E. (2009). *Fen ve teknoloji programı ilköğretim 1. ve 2. kademe öğretmen el kitabı* [Science and technology curriculum primary school teacher's handbook (grades 1–8)]. Pegem Akademi.
- Çetin, I., & Türkan, A. (2020). The impact of STEM activities on students' academic achievement: A meta-analytic review. *Education and Science*, 45(203), 55-72. <https://doi.org/10.15390/EB.2020.8460>
- Değerli, M. (2021). *Fen eğitiminde STEM yaklaşımının etkililiği: Bir meta analiz çalışması* [The effectiveness of the STEM approach in science education: A meta-analysis study] [Unpublished master's thesis]. Dicle University.
- Değerli, M., & Yapıcı, Ü. (2022). *Fen eğitiminde STEM yaklaşımının akademik başarıya etkisi: Bir meta-analiz çalışması* [The effect of the STEM approach in science education on academic achievement: A meta-analysis study]. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, 16(1), 17–48.
- Dweck, C.S. (2006). *Mindset: the new psychology of success*. random house.
- Ejiwale, J. A. (2013). Barriers to successful implementation of STEM education. *Journal of Education and Learning*, 7(2), 63-74.
- Ejiwale, J.A. (2013). Barriers to successful implementation of STEM education. *Journal of Education and Learning*, 7(2), 63-74.
- Fraenkel, J.R., & Wallen, N. E (2009). *How to design and evaluate research in education*. (7th Edition). McGraw-Hill Companies.
- Fraser, B.J. (1998). Classroom environment instruments: Development, validity, and applications. *Learning Environments Research*, 1(1), 7–33.
- Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3-8.
- Gonzalez, H.B., & Kuenzi, J.J. (2012, August). *Science, technology, engineering, and mathematics (STEM) education: A primer*. Congressional Research Service, Library of Congress.
- Hackman, S.T., Zhang, D., & He, J. (2021). Secondary school science teachers' attitudes towards STEM education in Liberia. *International Journal of Science Education*, 43(2), 223–246.
- Hedges, L.V., & Pigott, T.D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, 9(4), 426.
- Henriksen, D. (2014). Full STEAM ahead: Creativity in excellent STEM teaching practices. *The STEAM Journal*, 1(2), Article 15. <https://doi.org/10.5642/steam.20140102.15>

- Herdem, K., & Ünal, İ. (2018). Analysis of studies about STEM Education: A meta-synthesis study. *Journal of Educational Sciences*, 48(48), 145-163.
- Huedo-Medina T.B., Sanchez-Meca, J., Marin-Martinez F., & Botella, J. (2006). Assessing heterogeneity in meta analysis: Q statistics or I2 index? *Psychological Methods*, 11(2), 193-206.
- Kanadlı, S. (2019). *STEM eğitimi hakkında nitel bulguların meta-özeti* [A meta-summary of qualitative findings about STEM education]. *International Journal of Instruction*, 12(1), 959–976.
- Kaya, A. (2020). *Türkiye örnekleminde STEM eğitimi alanında yapılan çalışmaların içerik analizi* [Content analysis of studies conducted in Türkiye in the field of STEM education]. *İstanbul Aydın Üniversitesi Eğitim Fakültesi Dergisi*, 6(2), 275–306.
- Kazu, İ.Y., & Kaplan, A. (2024). *STEM uygulamalarının bilimsel süreç becerilerine etkisi: Bir meta-analiz çalışması* [The effect of STEM applications on scientific process skills: A meta-analysis study]. *Elektronik Sosyal Bilimler Dergisi*, 23(89), 234-255. <https://doi.org/10.17755/esosder.1331946>
- Kazu, İ.Y., & Kurtoglu Yalçın, C. (2021). The effect of STEM education on academic performance: a meta-analysis study. *Turkish Online Journal of Educational Technology-TOJET*, 20(4), 101-116.
- Kelley, T.R., & Knowles, J.G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, 3(1), 1–11. <https://doi.org/10.1186/s40594-016-0046-z>
- Kiazai, A.N., Siddiqua, N., & Waheed, Z. (2019). Challenges in implementing STEM education and role of teacher education programs in mitigating these challenges. *International Journal of Distance Education and E- Learning*, 1(1), 123-137.
- Kim, N.J., Belland, B.R., & Walker, A.E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychological Review* 30, 397–429.
- Landis, J.R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Lestari, T.P., Sarwi, S., & Sumarti, S.S. (2018). STEM-based project-based learning model to increase science process and creative thinking skills of 5th grade. *Journal of Primary Education*, 7(1), 18-24.
- Madden, L., Beyers, J., & O'Brien, S. (2016). The importance of STEM education in the elementary grades: Learning from pre-service and novice teachers' perspectives. *The Electronic Journal for Research in Science & Mathematics Education*, 20(5), 1–18.
- Marrero, M.E., Gunning, A.M., & Germain-Williams, T. (2014). What is STEM education? *Global Education Review*, 1(4), 1–6.
- Marshall, J.C., & Horton, R.M. (2011). The relationship of teacher-facilitated inquiry-based instruction to student higher-order thinking. *School Science and Mathematics*, 111(3), 93–101. <https://doi.org/10.1111/j.1949-8594.2010.00066.x>
- Martín-Páez, T., Aguilera, D., Perales-Palacios, F.J., & Vélchez-González, J.M. (2019). What are we talking about when we talk about STEM education? A review of literature. *Science Education*, 103(4), 799-822.
- McClain, M.L. (2015). *The effect of STEM education on mathematics achievement of fourth-grade underrepresented minority students* [Doctoral dissertation, Capella University]. ProQuest Dissertations & Theses Global.
- Miles, M.B., & Huberman, A.M. (2002). *The Qualitative Researcher's Companion*. Sage, CA.
- Mittlböck, M., & Heinzl, H.A. (2006). Simulation study comparing properties of heterogeneity measures in meta analyses. *Statistics in Medicine*, 25(24), 4321-4333.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: *The PRISMA Statement*. *BMJ*, 339, b2535.

- Mousoulides, N.G. (2013). Facilitating parental engagement in school mathematics and science through inquiry-based learning: An examination of teachers' and parents' beliefs. *ZDM*, 45(6), 863-874.
- National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. National Academies Press.
- Navruz, B., Erdogan, N., Bicer, A., Capraro, R.M., & Capraro, M.M. (2014). Would a STEM school 'by any other name smell as sweet'? *International Journal of Contemporary Educational Research*, 1(2), 67-75.
- Neely, J.G., Magit, A.E., Rich, J.T., Voelker, C.C., Wang, E.W., Paniello, R.C., & Bradley, J.P. (2010). A practical guide to understanding systematic reviews and meta-analyses. *Otolaryngology-Head and Neck Surgery*, 142(1), 6–14.
- Olaschinde, K.J., & Olatoye, R.A. (2014). Scientific attitude, attitude to science and science achievement of senior secondary school students in Katsina State, Nigeria. *Journal of Educational and Social Research*, 4(1), 445-452.
- Ormancı, Ü. (2020). Thematic content analysis of doctoral theses in STEM education: Turkey context. *Journal of Turkish Science Education*, 17(1), 126–146.
- Özkan, B., & Doğan, Y. (2022). The effect of STEAM activities on the academic achievement, scientific creativity and attitudes of seventh grade students. *Journal of Human Sciences*, 19(1), 37–54. <https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/5430>
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., & Mulrow, C.D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*;372(71). <http://doi.org/10.1136/bmj.n71>
- Parker, L.A., Saez, N.G., Porta, M., Herna'ndez-Aguado, I., & Lumbreras, B. (2013). The impact of including different study designs in meta-analyses of diagnostic accuracy studies. *European Journal of Epidemiology*, 28(9), 713-720.
- Pintrich, P.R., & De Groot, E.V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33–40.
- Pigott, T.D., & Polanin, J.R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, 90(1), 24–46.
- Quintana, S.M., & Minami, T. (2006). Guidelines for meta-analyses of counseling psychology research. *The Counseling Psychologist*, 34(6), 839-877.
- Saçan, E. (2018). *Stem-based curriculum proposal and effectiveness for science applications course*. [Unpublished master's thesis] Hacettepe University.
- Sanders, M. (2009). STEM, STEM education, STEMania. *The Technology Teacher*, 68(4), 20–27.
- Shelby, L.B., & Vaske, J.J. (2008). Understanding meta-analysis: a review of the methodological literature. *Leisure Sciences*, 30(2), 96-110.
- Strong, M.G. (2013). *Developing elementary math and science process skills through engineering design instruction*. [Unpublished master's thesis] Hofstra University.
- Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., & Song, F. (2000). *Methods for Meta-analysis in Medical Research*. John Wiley & Sons.
- Tarım, K. (2003). *Kübaşık öğrenme yönteminin matematik öğretimindeki etkinliği ve kübaşık öğrenme yöntemine ilişkin bir meta-analiz çalışması* [The effectiveness of the cooperative learning method in mathematics teaching and a meta-analysis study on the cooperative learning method] [Unpublished doctoral dissertation]. Çukurova University.
- Tate, W.F., Jones, B.D., Thorne-Wallington, E., & Hogrebe, M.C. (2012). Science and the city: Thinking geospatially about opportunity to learn. *Urban Education*, 47(2), 399–433.
- Tatlı, F. (2022). *Kimya öğretiminde STEM uygulamalarının farklı öğrenme stillerine ve zeka alanlarına sahip öğrencilerin kavramsal anlamaları ve bilimsel süreç becerilerine etkisi* [The effect of STEM applications in chemistry teaching on students with different learning

- styles and intelligence domains in terms of conceptual understanding and scientific process skills] [Unpublished doctoral dissertation]. Marmara University.
- Tolliver, E.R. (2016). *The effects of science, technology, engineering, and mathematics (STEM) education on elementary student achievement in urban schools* [Doctoral dissertation] Grand Canyon University.
- Toma, R.B., & Greca, I.M. (2018). The effect of integrative STEM instruction on elementary students' attitudes toward science. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(4), 1383-1395.
- Ulum, H. (2022). A meta-analysis of the effects of different integrated STEM (science, technology, engineering, and mathematics) approaches on primary students' attitudes. *International Journal of Educational Research Review*, 7(4), 307–317.
- Vennix, J., den Brok, P., & Taconis, R. (2018). Do outreach activities in secondary STEM education motivate students and improve their attitudes towards STEM? *International Journal of Science Education*, 40(11), 1263-1283.
- Wang, L.H., Chen, B., Hwang, G.J., Guan, J.Q., & Wang, Y.Q. (2022). Effects of digital game-based STEM education on students' learning achievement: a meta-analysis. *International Journal of STEM Education*, 9(1), 1-13.
- Xie, Y., Fang, M., & Shauman, K. (2015). STEM education. *Annual review of sociology*, 41, 331-357.
- Yakman, G. (2008). STΣ@ M Education: an overview of creating a model of integrative education. Pupils Attitudes Towards Technology. *2008 Annual Proceedings*. Netherlands.
- Yamak, H., Bulut, N., & Dündar, S. (2014). The impact of STEM Activities on 5th grade students' scientific process skills and their attitudes towards science. *GEFAD / GUJGEF* 34(2), 249-265.
- Yıldırım, B., & Altun, Y. (2015). Investigating the effect of STEM Education and engineering applications on science laboratory lectures. *El-Cezeri Journal of Science and Engineering*, 2(2), 28-40.
- Yıldırım, B. (2016). An analyses and meta-synthesis of research on STEM education. *Journal of Education and Practice*, 7(34), 23–33.
- Yıldırım, B. (2016). *An examination of the effects of science, technology, engineering, mathematics (STEM) applications, and mastery learning integrated into the 7<sup>th</sup> grade science course* [Unpublished master's thesis] Gazi University.
- Yılmaz, R.M., & Yılmaz, F.G.K. (2024). Gamified STEAM activities and their effects on student motivation and performance. *ERIC - Education Resources Information Center*. <https://files.eric.ed.gov/fulltext/EJ1437934.pdf>
- Yigit, M., & Bagci, H. (2024). Effects of STEAM education on students' creativity: A meta-analysis study. *Education Sciences*, 14(6), 676. <https://www.mdpi.com/2227-7102/14/6/676>



## APPENDIX

## Appendix A. Descriptive properties of the studies included in the meta-analysis

No	Title	Author(s)	N	Sample Characteristics
1	The effects of science technology-Engineering math (STEM) integration on 5th-grade students' perceptions and attitudes towards these areas	Gülhan and Şahin (2016)	57	5th-grade students
2	An examination of the effects of stem applications prepared in accordance with context-based learning	Yıldırım (2018)	26	Pre-service teachers
3	The Effects of STEM Activities on STEM Attitudes, Scientific Creativity and Motivation Beliefs of the Students and Their Views on STEM Education	Ugras (2018)	50	7th-grade students
4	Effect of STEM Activities on Students' Scientific Process Skills, Science Interest, Attitude, and Student Opinions	Simsek (2019)	52	7th-grade students
5	Investigation of the Effects of STEM Activities on Pre-Service Teachers' Self-Efficacy Beliefs and their STEM Intention Levels	Timur and Belek (2020)	104	Pre-service teachers
6	The Effect of Stem Applications on Students' Perceptions and Attitudes Towards Stem in The 6th Grade Science Course	Bahadır and Kose (2021)	73	6th-grade students
7	The Effects of Montessori Approach-Based STEM Activities on Pre-service Teachers' Attitudes Towards Science and Science Teaching	Cakir and Yalcin (2021)	100	Pre-service teachers
8	An Investigation of the Effects of STEM based Activities on Pre-service science Teachers' Science Process Skills	Gokbayrak and Karisan (2017)	50	Pre-service teachers
9	An Experimental Research on Effects of STEM Applications and Mastery Learning	Yildirim and Selvi (2017)	52	7th-grade students
10	The Effect of STEM Applications on 7th-Grade Students' Academic Achievement, Reflective Thinking Skills, and Motivations	Cakir and Ozan (2018)	53	7th-grade students
11	Teaching Applications' Based On 7E Learning Model Centered STEM Activity Effect on Academic Achievement	Guyen, Selvi and Benzer (2018)	37	5th-grade students
12	The Effects of STEM Training on the Academic Achievement of 4th Graders in Science and Mathematics and their Views on STEM Training Teachers	Acar, Tertemiz and Tasdemir (2018)	47	4th-grade students
13	The Impact of Teaching the Subject "Pressure" with STEM Approach on the Academic Achievements of the Secondary School 7th-Grade Students and Their Attitudes Towards STEM	Özcan and Koca (2019)	33	7th-grade students
14	The Effect of Stem Activities on Pre-school Students' Scientific Process Skills	Keçeci, Aydın and Zengin (2019)	24	Pre-school students
15	The Effect of STEM Activities on Students' Achievement in "Sound" Subject	Dedetürk, Kırmızıgül and Kaya (2019)	158	6th-grade students
16	An Investigation the Effect of STEM Practices on Fifth Grade Students' Academic Achievement and Motivations at The Unit "Exploring and Knowing the World of Living Creatures"	Parlakay and Koç (2020)	64	5th-grade students
17	The Effect of STEM Implementation on Attitude Towards Stem And Success in "Measurement of Force and Friction" Class	Ozan and Sagir (2020)	20	5th-grade students
18	The Effects of STEM Activities on 8th-Grade Students' Science Process Skills, Scientific Epistemological Beliefs, and Science Achievements	Bahsi and Fırat (2020)	32	8th-grade students
19	The Effect of Stem Applications on the Stem Awareness of Students and the Performance of the Success in the "Triangles" Unit	Gürbüz and Karadeniz (2020)	33	9th-grade students
20	The Effect of The Stem Approach Based on the 5E Model on Academic Achievement and Scientific Process Skills: The Transformation of Electrical Energy	İzgi and Kalaycı (2020)	50	7th-grade students
21	The Effect of STEM Activities Prepared According to the Design Thinking Model on Pre-school Children's Creativity and Problem-Solving Skills	Yalçın and Erden (2021)	39	Pre-school students
22	The Effect of STEM-based Education Program on Problem-Solving Skills of Five-Year-old Children	Şahin (2021)	37	Pre-school students



## Investigating homogeneity of variance in normal, skewed-normal, and gamma distributions: A simulation study

Serpil Çelikten-Demirel<sup>1\*</sup>, Ayşenur Erdemir<sup>2</sup>, Esra Oyar<sup>3</sup>, Tuba Gündüz<sup>4</sup>

<sup>1</sup>Dicle University, Ziya Gokalp Faculty of Education, Department of Educational Sciences, Diyarbakır, Türkiye

<sup>2</sup>Turkish National Police Academy, Institute of Forensic Sciences, Department of Forensic Psychology, Ankara, Türkiye

<sup>3</sup>Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

<sup>4</sup>Mugla Sitki Kocman University, Faculty of Education, Department of Educational Sciences, Muğla, Türkiye

### ARTICLE HISTORY

Received: Dec. 31, 2024

Accepted: Aug. 28, 2025

### Keywords:

Homogeneity of variances,  
Bartlett's test,  
Levene's test,  
Brown-Forsythe test,  
Fligner-Killeen test.

**Abstract:** It is an important point to test the homogeneity of variances in statistical methods such as the *t*-test or *F*-test used to make comparisons between groups. An erroneous decision regarding the homogeneity of variances will affect the test to be selected and thus lead to different results. For this reason, there are many tests for homogeneity of variance in the literature. This study aims to examine the type I error and power ratios of Levene, Bartlett, Brown-Forsythe, and Fligner-Killeen tests under different conditions. In this study, conducted within the scope of basic research, analyses were performed using simulated data. The simulation conditions included variance ratio (1:1, 1:2, 1:3, 2:1, 3:1), distributions (normal, skewed-normal, gamma), sample sizes (60, 120, and 240), and ratio of group sizes (1/1, 1/2, 1/4, 1/9). According to the study results, when controlling for type I error is a primary concern, the Brown-Forsythe and Fligner-Killeen tests are recommended, particularly under non-normal distributions. If the power is a major concern for research, the Bartlett's test and the Levene's test should be used in general.

## 1. INTRODUCTION

In many studies conducted in social sciences, various demographic variables are discussed, and inferences are made by comparing group averages according to these variables. To draw these inferences, statistical tests are employed. The choice of test depends on the characteristics of the data and is generally classified as either parametric or nonparametric, based on whether the relevant assumptions are met. Non-parametric tests are easier to calculate than parametric tests. However, they are less powerful, being less likely to reject a false null hypothesis (Woodbury, 2002, p. 591). Consequently, when the assumptions for parametric tests are satisfied, their use is recommended. These assumptions primarily involve normality and homogeneity of variances, which are fundamental when comparing group means in parametric statistics (Tabachnick & Fidell, 2007, p. 201).

\*CONTACT: Serpil ÇELİKTEN-DEMİREL ✉ [serpil.celikten@dicle.edu.tr](mailto:serpil.celikten@dicle.edu.tr) 📍 Dicle University, Faculty of Education, Department of Educational Sciences, Diyarbakır, Türkiye

The copyright of the published article belongs to its author under CC BY 4.0 license. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

Because many variables in science and nature follow a normal distribution, it is the most widely applied distribution in statistics (Kirk, 2008, p.230). Most statistical methods, including correlation, regression, and group comparisons, including the *t*-test and ANOVA *F*-test, are grounded in the normal distribution and rely on the assumption of normality. In this case, it is necessary to examine whether the data have a normal distribution before using the relevant tests (Orcan, 2020; Sedgwick, 2015; Woodbury, 2002). At this point, the tests used to test the normality assumption under different simulation conditions are compared, and evaluations are made on the strength of the test. Another key assumption is homogeneity of variance, which means that the variance of a variable remains constant across the levels of another variable (Howell, 2010, p.213). Box (1954) stated that the *F*-test is robust to violations of the homogeneity of variances assumption provided that (1) there are equal numbers of observations at each variable level, (2) the population distributions are normal, and (3) the ratio of the largest variance to the smallest variance does not exceed 3. However, studies have shown that even when sample sizes are equal, the *F*-test is not robust against heterogeneity of variances, which is frequently encountered in social and educational sciences research. In this case, it is important that researchers should not ignore violations of the homogeneity of variances assumption (Kirk, 2008, pp.411-412). In addition, if the variances of the groups are not homogeneous, the inferences to be obtained as a result of the use of parametric tests may not be valid. In addition, although the variances are homogeneous, if this cannot be detected, invalid inferences may be obtained as a result of the use of nonparametric tests due to lower statistical power.

Various tests have been developed to examine the homogeneity of variances. The null hypothesis for all tests considered in this study is that variances are equal between groups and is shown as follows:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_k^2$$

On the other hand, the case that the variances of at least two populations are not equal to each other is established as the alternative hypothesis and is shown as follows:

$$H_1: \sigma_i^2 = \sigma_j^2 \quad \text{for some } 1 \leq i \neq j \leq k$$

One of the frequently used tests for testing homogeneity of variance is Levene's Test for Equality of Variance (Field, 2018, p.346; Gamst *et al.*, 2008, p.58). Since Levene's test is not calculated depending on the sample variance, it is not affected by outliers (Zhou *et al.*, 2023). However, it is stated that the power of Levene's test to detect variance differences between levels of a variable depends on the amount of data collected. In large samples, small differences in variances lead to a significant Levene's test, while conversely, in small samples, relatively large differences in variances may not be detected. In other words, in large samples, small differences in variance may increase the significance of Levene's test. However, in small sample conditions, even quite large differences may not be detected (Field, 2018, p.359). In this case, sample size is a point to be considered when using Levene's test.

Another method used to test the homogeneity of variances is Bartlett's test (Bartlett, 1937). Similar to Levene's test, it tests the null hypothesis that the variances in the population are equal. Since Bartlett's test is derived from the likelihood ratio test under normal distribution, it is affected by the normality assumption (Arsham & Lovric, 2011). Therefore, it is stated that this test is powerful when populations are normally distributed (Arsham & Lovric, 2011; Glass, 1966). In cases where populations are not normally distributed, Bartlett's test can be used if the sample is large enough. On the other hand, Levene's test is stated as a powerful test for situations where normal distribution is not provided (Gastwirth *et al.*, 2009).

Another method used to test the homogeneity of variances is the Fligner-Killeen test (Fligner & Killeen, 1976). It is stated that it is a nonparametric test for testing the homogeneity of variance between groups. Also, it can be used in cases where the data do not show a normal distribution. A robust statistical method for determining if variances in groups are equal, the

Brown-Forsythe test was created specifically to overcome Levene's test's shortcomings when dealing with non-normal distributions. It is a modified form of Levene's test where the absolute deviations of observations within each group are calculated using the median (or other trimmed means) in place of the mean (Brown & Forsythe, 1974).

Many studies are comparing the performance of the tests used to test the homogeneity of variances under different conditions (Abdullah & Muda, 2022; Conover *et al.*, 1981; Katsileros *et al.*, 2024; Keskin, 2002; Kim & Cribbie, 2017; Öztürk, 2020; Park, 2018; Wang *et al.*, 2017; Yonar *et al.*, 2024). While some of these studies analyzed only simulation data, others analyzed both simulation and real data. Within the scope of simulation studies, percent of correct rejection (power) and percent of false rejection (type I error rate) values were compared for sample size (Abdullah & Muda, 2022; Gökpınar, 2020; Kim & Cribbee, 2017; Wang *et al.*, 2017; Yonar, 2024), number of groups (Gökpınar, 2020; Kim & Cribbee, 2017; Zhou *et al.*, 2023), variance ratios between groups (Kim & Cribbee, 2017; Wang *et al.*, 2017; Zhou *et al.*, 2023) and different distribution types (such as normal distribution, Laplace distribution, Chi-square distribution; Wang *et al.*, 2017; Yonar *et al.*, 2024; Zhou *et al.*, 2023).

This study aims to examine various tests for homogeneity of variances under different distributions, different variance ratios between groups, and different sample sizes, and sample ratios, and to propose the most appropriate test for various conditions likely to be encountered in practice by crossing possible conditions. Gamma, skewed-normal, and normal distributions were utilized for the distribution condition. Total sample sizes of 60, 120, and 240 were considered, along with group size ratios of 1:1, 1:2, 1:4, and 1:9, and variance ratios of 1:1, 1:2, 1:3, 2:1, and 3:1. The performance of Bartlett's test, Levene's test, Fligner-Killeen test, and the Brown-Forsythe test was evaluated under these conditions.

In this direction, the research questions of the study are as follows:

1. What are the error rates (false rejections) of the tests for homogeneity of variances for varying sample sizes and sample ratios under different distribution conditions when the variance condition is 1:1?
2. What are the power ratios (correct rejections) of the tests for homogeneity of variances under different sample sizes, sample ratios, and variance ratios for normal and skewed-normal distributions?

What are the power ratios (correct rejections) of the tests for homogeneity of variances under different sample sizes, sample ratios, and variance ratios for gamma distributions?

## 2. METHOD

### 2.1. Study Design

Within the scope of the study, the power and type I error rates of the tests developed to test the homogeneity of variances under different conditions were evaluated. Basic research includes studies aimed at revealing the processes underlying a theoretical hypothesis (Fraenkel & Wallen, 2009, p.7). In this context, the study is basic research because there are hypotheses about the tests discussed in the study, and the validity of these hypotheses is tested under different conditions.

### 2.2. Data

Within the scope of the study, data were generated over the beta distribution, skewed-normal, and gamma distribution. The data were generated in a Python environment. Descriptive statistics were analyzed to check the validity of the generated data. The data were generated in such a way that there are two groups within each data set. The variables simulated in the study are as follows:

### 2.2.1. Ratio of variance between groups

In group comparison, the homogeneity of variances is expressed as the variances of the groups being equal to each other (Kirk, 2008, p.326). Since it is not very common for the variance ratios of the groups to be greater than 3 (Kirk, 2008), the variance ratios between the pairs were set at 1:1, 1:2, and 1:3. In addition, 2:1 and 3:1 ratios were included to evaluate conditions of small sample-large variance and large sample-small variance. Thus, five levels were considered for the variance ratio condition.

### 2.2.2. Data distribution

The type of distribution affects the significance of homogeneity tests (Brown & Forsythe, 1974; Gastwirth *et al.*, 2009). Given its frequent use in the social sciences, normal distribution plays a critical role in statistical tests (Shavelson, 1996, p. 115). In addition, skewed-normal distribution, which is close to normal distribution, has also been used in studies on statistical tests (Arnold *et al.*, 2014; Sarısoy *et al.*, 2013; Zhou *et al.*, 2023). In addition to these distributions, a systematic review study conducted by Bono and colleagues (2017) examined the types of distributions used in studies in the fields of health, education, and social sciences. Among the 262 studies reviewed in the Web of Science database between 2010 and 2015, the gamma distribution emerged as the most frequently used non-normal distribution, appearing in 57 articles. As a result, three different distribution types were used in the study: normal, skewed-normal, and gamma distribution. For normal distribution,  $N \sim (0, 1, 0)$ , for skewed-normal distribution, slightly skewed (shape=2) and highly skewed (shape=10), and for gamma distribution, slightly skewed (shape=5, scale=1) and highly skewed (shape=2, scale=1) (Ahsanullah, 2017, p. 24; Azzalini, 1985, p. 174).

### 2.2.3. Sample sizes and ratio of group sizes

In studies conducted in the field of social sciences, sample size is a point to be considered in order to avoid errors in interpreting the results. In studies with different subgroups and comparisons between groups (e.g., gender, school, marital status, etc.), it is recommended to have at least 30 units from each subgroup; in experimental studies, it is recommended to have a sample size between 10-20 (Roscoe, 1975). Four different sample sizes ( $N = (n_1 + n_2)$ ): 60, 120, and 240, and four levels of ratio of group sizes ( $n_1/n_2 = 1/1, 1/2, 1/4, 1/9$ ) were investigated for each distribution.

As a result, a total of 300 different conditions were obtained, 5 (variance ratios) x 5 (distributions) x 3 (sample sizes) x 4 (ratio of group sizes). For each condition, 10000 replications were generated.

## 2.3. Data Analysis

Three different variance homogeneity tests, which are Bartlett's test, Levene's test, Fligner-Killeen test, and the Brown-Forsythe test, were considered in the analysis of the data. All the analysis was done in a Python environment. Then, the type I error (false rejection) and power (correct rejection) rates were compared over the test results. When the variances were equal, the performance of the tests was evaluated according to the false rejection (type I error). On the other hand, when the variances were not equal, the performance of the tests was evaluated according to the correct rejection (power).

## 3. RESULTS

The simulation study was conducted to compare the empirical Type I error rates and the statistical power of the four different homogeneity tests of variances, manipulating the type of distributions, sample sizes, ratio of group sizes, and ratio of group variances. Results were examined based on the crossed conditions mentioned in the methods section.

Three different sample sizes ( $N=(n_1+n_2)$ ): 60, 120, 240, and four levels of ratio of group sizes ( $n_1/n_2=1/1; 1/2; 1/5; 1/9$ ) were investigated under the normal, the skewed-normal distributions

(slightly-skewed [ $\alpha=2$ ], highly-skewed [ $\alpha=10$ ]), and the gamma distributions (high degree of skewness-GA [2,1]; slight degree of skewness-GA [5,1]), respectively.

### 3.1. Results of the Type I Error Rates under Each Distribution

In this section, the false rejection rates (Type I error) of the Bartlett, Levene, Brown-Forsythe, and Fligner-Killeen tests were evaluated under normal, skewed-normal, and gamma distributions when group variance ratios were equal. These examinations were conducted under a range of conditions where the ratios of group sizes were 1/1, 1/2, 1/5, 1/9, and the total sample size was 60, 120, and 240 in each condition, respectively. Table 1 summarizes the error rates (false rejections/type I error) for the results of the tests under each distribution.

**Table 1.** The empirical type I error rates under normal, skewed-normal, and gamma distributions in terms of different ratios of group sizes and sample sizes.

N	RoGS ( $n_1, n_2$ )		Normal Distribution	Skewed-Normal Distribution		Gamma Distribution	
				Slightly- skewed	Highly- skewed	Slightly- skewed	Highly- skewed
60	1/1 (30, 30)	BRT	0.051	0.070	0.098	0.104	0.180
		LEV	0.052	0.066	0.093	0.071	0.102
		BF	0.041	0.048	0.049	<b>0.041</b>	0.049
		FK	<b>0.040</b>	0.047	0.055	0.045	0.061
	1/2 (20, 40)	BRT	0.049	0.068	0.093	0.103	0.177
		LEV	0.051	0.059	0.091	0.072	0.104
		BF	0.042	0.044	0.047	0.043	0.047
		FK	0.042	0.045	0.057	0.047	0.056
	1/5 (10, 50)	BRT	0.048	0.065	0.082	0.094	0.160
		LEV	0.053	0.061	0.080	0.070	0.101
		BF	0.045	0.046	0.044	0.042	0.044
		FK	0.044	0.046	0.055	0.048	0.062
	1/9 (6, 54)	BRT	0.049	0.059	0.076	0.081	0.127
		LEV	<b>0.054</b>	0.055	0.077	0.068	0.089
		BF	0.047	<b>0.040</b>	<b>0.040</b>	<b>0.041</b>	<b>0.039</b>
		FK	0.048	0.046	0.055	0.049	0.062
120	1/1 (60, 60)	BRT	0.053	0.065	0.096	0.112	0.195
		LEV	0.052	0.058	0.088	0.072	0.099
		BF	0.047	0.045	0.049	0.045	0.049
		FK	0.045	0.046	0.060	0.050	0.062
	1/2 (40, 80)	BRT	0.050	0.070	0.100	0.105	0.189
		LEV	0.052	0.060	0.088	0.071	0.097
		BF	0.046	0.048	0.051	0.047	0.046
		FK	0.044	0.048	0.062	0.049	0.062
	1/5 (20, 100)	BRT	0.050	0.069	0.093	0.099	0.177
		LEV	0.050	0.060	0.088	0.068	0.096
		BF	0.046	0.048	0.052	0.044	0.046
		FK	0.044	0.049	0.062	0.050	0.062
	1/9 (12, 108)	BRT	0.049	0.061	0.085	0.093	0.163
		LEV	0.051	0.056	0.082	0.063	0.095
		BF	0.045	0.046	0.045	0.042	0.043
		FK	0.046	0.046	0.056	0.050	0.059



**Table 1.** Continued.

N	RoGS ( $n_1, n_2$ )		Normal Distribution	Skewed-Normal Distribution		Gamma Distribution	
				Slightly- skewed	Highly- skewed	Slightly- skewed	Highly- skewed
240	1/1 (120, 120)	BRT	0.049	<b>0.072</b>	<b>0.101</b>	<b>0.117</b>	<b>0.201</b>
		LEV	0.049	0.063	0.088	0.072	0.096
		BF	0.046	0.054	0.050	0.050	0.044
		FK	0.046	0.054	0.064	0.054	0.063
	1/2 (80, 160)	BRT	0.047	0.067	0.097	0.116	0.195
		LEV	0.048	0.057	0.082	0.071	0.096
		BF	0.045	0.049	0.047	0.050	0.050
		FK	0.045	0.050	0.060	0.054	0.069
	1/5 (40, 200)	BRT	0.052	0.067	0.095	0.110	0.191
		LEV	0.051	0.058	0.086	0.073	0.099
		BF	0.049	0.048	0.051	0.048	0.047
		FK	0.048	0.049	0.063	0.052	0.066
	1/9 (24, 216)	BRT	0.050	0.067	0.094	0.105	0.184
		LEV	0.050	0.057	0.091	0.069	0.097
		BF	0.048	0.046	0.054	0.044	0.045
		FK	0.048	0.048	0.068	0.050	0.066

N: Total sample size; RoGS: Ratio of group size; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Error ratios in Table 1 show that the Fligner-Killeen test and the Brown-Forsythe test produced the lowest type I error rates with values close to each other for each crossed condition in terms of distribution, sample size, and group size (e.g., under the normal distribution with  $N=60$  and RoGS: 1/1); however, the results varied, at times favoring the Fligner-Killeen and, alternatively, the Brown-Forsythe test. Detailed examinations of the lowest and highest error rates for tests of homogeneity of variances are presented for each of the distributions, considering the group ratio sizes.

For normal distribution, the Fligner-Killeen test yielded the best result with the lowest type I error rate for the balanced group sizes of 1/1 (30, 30). The highest type I error rate was observed in Levene's test with the group ratio of 1/9 (6, 54). For the slightly-skewed normal distribution, the Brown-Forsythe test showed the lowest type I error rate with the group ratio of 1/9 (6, 54). The highest error rates were produced by the BRT test for the balanced group sizes of 1/1 (120, 120). For the highly-skewed normal distribution, the Brown-Forsythe test again yielded the lowest type I error rate with the group ratio of 1/9 (6, 54), and the highest error rates came from Bartlett's test for the group sizes of 1/1 (120, 120), as in the slightly-skewed normal distribution.

For the slightly-skewed gamma distribution, the Brown-Forsythe test yielded the lowest error rate for the group ratio sizes of 1/1 (30, 30) and 1/9 (6, 54), respectively. The highest type I error rate was observed in the Bartlett's test with the group ratio sizes of 1/1 (120, 120). For the highly-skewed gamma distribution, the Brown-Forsythe test again produced the lowest error rate with the group ratio sizes of 1/9 (6, 54), while Bartlett's test produced the highest error rate for the group size ratio of 1/1 (120, 120).

When considering the varying degrees of skewness in both the skewed-normal and gamma distributions simultaneously, it was observed that methods producing the highest and lowest error rates were consistent across distributions. The Brown-Forsythe test resulted in the lowest error rates with the ratio of group sizes of 1/9 (6, 54), while Bartlett's test showed the highest error rates for the balanced group sizes of 1/1 (120, 120). An exception occurred in the slightly

skewed gamma distribution, where the Brown-Forsythe test also produced the lowest error rate for the group ratio of 1/1 (30, 30), equal to the lowest error observed for the group ratio of 1/9 (6, 54). However, the cases yielding the highest and lowest error rates under the normal distribution occurred under different conditions than those observed in the skewed-normal and gamma distributions.

In order to achieve a more profound comprehension of the outcomes, it is crucial to undertake a holistic analysis of the conditions under the normal, skewed-normal, and gamma distributions. Analyses revealed that the error values observed under the gamma distribution generally tended to be higher than those observed under the normal and skewed-normal distributions.

When analyzing each distribution separately, it was found that under conditions where all sample sizes and group proportions were crossed, the highest type I error rates occurred in Bartlett's test when both distributions were highly skewed, except for one case arising from Levene's test. In some instances, Levene's test resulted in higher error rates than Bartlett's test, while in others, Bartlett's test produced higher errors than Levene's. A closer examination of these discrepancies showed that the results of the two tests were highly similar.

A comparison of the less skewed and highly skewed distributions, under both the skewed-normal and gamma conditions, showed that the highly skewed distribution produced higher error rates. Methodologically, the lowest error values were obtained with the Brown-Forsythe and Fligner-Killeen tests, though the difference between them was relatively small.

An evaluation of the results in terms of group size ratios, with each sample size held constant, revealed that differences in group sizes generally led to a decline in type I error, though a few cases contradicted this pattern. Lastly, when examining the impact of sample size on each method while maintaining constant group sizes across distributions, it was found that the observed fluctuations were not systematic.

### **3.2. Results of the Power for the Normal and Skewed-Normal Distributions**

The powers of tests for homogeneity of variance were computed for skewed-normal distributions, including the normal, slightly skewed, and highly skewed normal distributions, separately. These computations were conducted under four different group variance ratio conditions (1:2, 1:3, 2:1, 3:1) and across different sample sizes with varying group size ratios (1/1, 1/2, 1/5, 1/9). Results are presented in [Table 2](#), with the highest and lowest values highlighted in bold for each distribution crossed by the variance ratio conditions.

**Table 1.** Correct rejection rate under normal and skewed-normal distribution with varying conditions in terms of sample size, ratio of group size, and ratio of group variance.

N	RoGS ( $n_1, n_2$ )	VR HT	Normal distribution				Slightly-skewed- normal distribution				Highly-skewed- normal distribution			
			1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
60	1/1 (30, 30)	BRT	0.441	0.826	0.459	0.831	0.450	0.812	0.453	0.811	0.460	0.792	0.457	0.787
		LEV	0.390	0.758	0.398	0.763	0.387	0.737	0.396	0.741	0.412	0.738	0.416	0.743
		BF	0.352	0.722	0.363	0.728	0.340	0.688	0.346	0.699	0.308	0.644	0.314	0.649
		FK	0.327	0.684	0.340	0.690	0.319	0.654	0.331	0.665	0.314	0.632	0.313	0.632
	1/2 (20, 40)	BRT	0.379	0.770	0.422	0.788	0.399	0.753	0.419	0.764	0.409	0.737	0.421	0.743
		LEV	0.330	0.678	0.383	0.731	0.338	0.657	0.376	0.706	0.357	0.673	0.395	0.708
		BF	0.302	0.653	0.334	0.692	0.300	0.617	0.320	0.653	0.262	0.565	0.295	0.615
		FK	0.295	0.633	0.298	0.629	0.300	0.612	0.286	0.588	0.290	0.593	0.272	0.561
	1/5 (10, 50)	BRT	0.221	0.486	0.283	0.584	0.240	0.500	0.282	0.569	0.266	0.518	0.291	0.547
		LEV	0.188	0.407	0.278	0.554	0.185	0.394	0.274	0.540	0.230	0.437	0.294	0.540
		BF	0.178	0.393	0.227	0.489	0.163	0.360	0.221	0.472	0.147	0.320	0.205	0.437
		FK	0.188	0.407	0.188	0.401	0.181	0.393	0.173	0.384	0.194	0.407	0.168	0.354
	1/9 (6, 54)	BRT	0.139	0.273	0.195	0.408	0.155	0.293	0.200	0.393	0.174	0.317	0.203	0.389
		LEV	0.114	0.219	0.210	0.419	0.111	0.212	0.211	0.401	0.151	0.261	0.226	0.407
		BF	<b>0.110</b>	<b>0.212</b>	0.167	0.353	<b>0.098</b>	<b>0.187</b>	0.161	0.332	<b>0.086</b>	<b>0.166</b>	0.149	0.308
		FK	0.120	0.232	<b>0.134</b>	<b>0.266</b>	0.122	0.228	<b>0.120</b>	<b>0.247</b>	0.140	0.247	<b>0.118</b>	<b>0.229</b>
120	1/1 (60, 60)	BRT	0.749	0.987	0.752	0.985	0.734	0.980	0.741	0.982	0.720	0.969	0.725	0.969
		LEV	0.690	0.973	0.686	0.970	0.664	0.964	0.674	0.964	0.669	0.957	0.684	0.956
		BF	0.672	0.969	0.669	0.967	0.634	0.957	0.646	0.959	0.587	0.934	0.600	0.934
		FK	0.647	0.960	0.644	0.955	0.612	0.944	0.618	0.944	0.582	0.924	0.598	0.927
	1/2 (40, 80)	BRT	0.683	0.978	0.705	0.973	0.677	0.967	0.698	0.964	0.668	0.952	0.670	0.949
		LEV	0.619	0.949	0.649	0.952	0.595	0.937	0.637	0.943	0.611	0.931	0.637	0.929
		BF	0.606	0.946	0.624	0.943	0.573	0.930	0.598	0.932	0.519	0.903	0.555	0.905
		FK	0.592	0.936	0.590	0.922	0.561	0.921	0.561	0.907	0.540	0.903	0.534	0.878
	1/5 (20, 100)	BRT	0.458	0.856	0.507	0.866	0.465	0.842	0.500	0.845	0.475	0.829	0.498	0.820
		LEV	0.387	0.781	0.472	0.835	0.389	0.753	0.468	0.806	0.414	0.763	0.481	0.800
		BF	0.388	0.781	0.430	0.809	0.370	0.738	0.419	0.774	0.325	0.682	0.393	0.741
		FK	0.388	0.778	0.389	0.747	0.378	0.745	0.369	0.709	0.374	0.727	0.349	0.675
	1/9 (12, 108)	BRT	0.288	0.623	0.356	0.693	0.301	0.621	0.361	0.681	0.325	0.627	0.357	0.658
		LEV	0.241	0.532	0.342	0.668	0.236	0.511	0.350	0.653	0.272	0.551	0.355	0.651
		BF	0.248	0.542	0.297	0.626	0.224	0.497	0.297	0.604	0.196	0.439	0.271	0.565
		FK	0.257	0.548	0.254	0.547	0.246	0.526	0.249	0.517	0.257	0.533	0.228	0.471
240	1/1 (120, 120)	BRT	<b>0.963</b>	<b>1.000</b>	<b>0.967</b>	<b>1.000</b>	<b>0.952</b>	<b>1.000</b>	<b>0.950</b>	<b>1.000</b>	<b>0.937</b>	<b>1.000</b>	<b>0.933</b>	<b>1.000</b>
		LEV	0.934	<b>1.000</b>	0.939	<b>1.000</b>	0.927	0.999	0.920	<b>1.000</b>	0.918	0.999	0.915	0.999
		BF	0.932	<b>1.000</b>	0.936	<b>1.000</b>	0.919	<b>1.000</b>	0.912	<b>1.000</b>	0.887	0.999	0.885	0.999
		FK	0.921	<b>1.000</b>	0.927	<b>1.000</b>	0.905	0.999	0.898	<b>1.000</b>	0.883	0.998	0.881	0.998
	1/2 (80, 160)	BRT	0.944	<b>1.000</b>	0.946	<b>1.000</b>	0.934	<b>1.000</b>	0.927	0.999	0.914	<b>1.000</b>	0.905	0.999
		LEV	0.907	<b>1.000</b>	0.915	0.999	0.891	0.999	0.897	0.999	0.886	0.999	0.883	0.998
		BF	0.905	<b>1.000</b>	0.907	0.999	0.882	0.999	0.886	0.998	0.847	0.998	0.845	0.997
		FK	0.898	0.999	0.894	0.998	0.872	0.998	0.863	0.997	0.853	0.998	0.832	0.995

N: Total sample size; RoGS: Ratio of group size; HT: Homogeneity tests; VR: Variance ratio; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

**Table 2.** *Continued.*

N	RoGS (n <sub>1</sub> , n <sub>2</sub> )	VR HT	Normal distribution				Slightly-skewed- normal distribution				Highly-skewed- normal distribution			
			1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
240	1/5 (40, 200)	BRT	0.776	0.996	0.797	0.990	0.774	0.990	0.782	0.988	0.758	0.985	0.748	0.976
		LEV	0.713	0.986	0.760	0.984	0.692	0.979	0.739	0.976	0.698	0.973	0.722	0.968
		BF	0.716	0.986	0.740	0.981	0.682	0.979	0.713	0.972	0.625	0.959	0.661	0.957
		FK	0.711	0.984	0.705	0.970	0.682	0.978	0.670	0.959	0.660	0.965	0.628	0.936
	1/9 (24, 216)	BRT	0.564	0.940	0.624	0.935	0.572	0.926	0.604	0.919	0.568	0.909	0.587	0.895
		LEV	0.492	0.894	0.593	0.913	0.476	0.874	0.569	0.898	0.503	0.869	0.575	0.879
		BF	0.504	0.900	0.560	0.900	0.471	0.871	0.528	0.878	0.419	0.817	0.498	0.843
		FK	0.505	0.898	0.519	0.871	0.481	0.875	0.480	0.840	0.479	0.852	0.453	0.795

N: Total sample size; RoGS: Ratio of group size; HT: Homogeneity tests; VR: Variance ratio; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Results in Table 2 indicated that Bartlett's test tended to provide the best power among the homogeneity of variances tests considered in this study, apart from a few cases where Levene's test showed slightly higher power, though with only a small increase.

Under the normal distribution, all homogeneity of variance tests achieved perfect power with variance ratios of 1:3 and 3:1 for the balanced group sizes of 1/1 (120, 120). For the group size ratio of 1/2 (80, 160), Bartlett's, Levene's, and Brown-Forsythe tests yielded perfect power with a variance ratio of 1:3, while for the variance ratio of 3:1, only Bartlett's test achieved perfect power. For the variance ratios of 1:2 and 2:1, again with the balanced group sizes of 1/1 (120, 120), Bartlett's test provided the highest percentage of correct rejections. Results under the slightly skewed normal distribution indicated that, for the balanced group size of 1/1 (120, 120) with a variance ratio of 3:1, all methods produced perfect power. For the group size ratio of 1/2 (80, 160), Bartlett's and the Brown-Forsythe test achieved perfect power under the variance ratio of 1:3. Moreover, for the 1:2 and 2:1 variance ratios with the balanced group sizes of 1/1 (120, 120), Bartlett's test again provided the highest power, consistent with the normal distribution results. Under the highly skewed normal distribution, results showed that for the balanced group size of 1/1 (120, 120) with a variance ratio of 3:1, and for the group size ratio of 1/2 (80, 160), Bartlett's test achieved perfect power in each case. For the variance ratios of 1:2 and 2:1, Bartlett's test again showed the highest power for the balanced group size of 1/1 (120, 120). Examining the lowest true rejection rates, the Brown-Forsythe test produced the lowest power for variance ratios of 1:2 and 1:3 with the group size ratio of 1/9 (6, 54) under the normal, slightly skewed, and highly skewed normal distributions, respectively. For the variance ratios of 2:1 and 3:1 with the group size ratio of 1/9 (6, 54), the Fligner-Killeen test resulted in the lowest power under the normal, slightly skewed, and highly skewed normal distributions in the same manner.

When group sizes remained constant, it was observed that an increase in the overall sample size systematically led to higher power of the tests across all distribution types. Moreover, when the total sample size remained constant, increasing the disparity between group sizes reduced statistical power. In summary, larger sample sizes consistently increased the power of the tests, whereas greater differences in group size ratios decreased it.

In general, Bartlett's test tended to provide the highest power rates compared to the other methods under all distributions crossed by variance ratio conditions. Only under the group size ratio of 1/9 (6, 54) did Levene's test produce higher power values than Bartlett's test in some cases (e.g., under the normal distribution with group variance ratios of 2:1 and 3:1, and under the slightly skewed distribution with group variance ratios of 2:1 and 3:1). These differences were small, with a maximum of 0.018, suggesting that they may have been random rather than

systematic. When the lowest correct rejection rates were examined, the Brown-Forsythe and Fligner-Killeen tests tended to produce the lowest values under all conditions, although in some cases their values were the same or very similar.

Increasing the group variance ratios led to a higher correct rejection rate of the homogeneity of variance tests (the expected result). However, changing the order of the group variances (e.g., 2:1 instead of 1:2) did not significantly impact the correct rejection rates. The small changes observed between the group variance conditions of 1:2 and 2:1, and between 1:3 and 3:1 under each distribution, were not systematic. When the differences between the 1:2–2:1 and 1:3–3:1 under each distribution condition (for example, the difference of 1:2 and 2:1 under normal distribution) were analyzed, it was observed that these difference values were generally low. When these differences were analyzed quantitatively, it was found that only 6% of the values were above 0.1, with the remaining values being 0.1 or below. However, this variation was not systematic.

In circumstances where the total sample size is minimal and the rate of differentiation between group sizes is substantial, it is imperative to meticulously select the most suitable method. In such cases, although Bartlett's test generally stood out, both Bartlett's and Levene's tests gained particular importance based on the quantitative comparison of correct rejection rates. It was established that the efficacy of the methods is directly proportional to the total sample size, with power values approaching 1. However, it was also demonstrated that an increase in the discrepancy between group sizes adversely affected the observed power values.

### 3.3. Results of the Power for the Gamma Distributions

**Table 3** summarizes the correct rejection rates under the gamma distributions across varying conditions of sample size, ratio of group size, and ratio of group variance. The highest and lowest results are highlighted in bold for each distribution crossed by variance ratio conditions.

**Table 2.** Correct rejection rate under gamma distribution across varying conditions in terms of sample size, ratio of group size, and ratio of group variance.

N	RoGS (n <sub>1</sub> ,n <sub>2</sub> )	VR HT	Slightly-skewed- gamma distribution				Highly-skewed-gamma distribu- tion			
			1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
60	1/1 (30, 30)	BRT	0.463	0.792	0.456	0.786	0.468	0.740	0.468	0.744
		LEV	0.382	0.720	0.383	0.724	0.366	0.654	0.374	0.666
		BF	0.307	0.651	0.308	0.651	0.248	0.538	0.254	0.542
		FK	0.305	0.629	0.306	0.638	0.279	0.562	0.285	0.570
	1/2 (20, 40)	BRT	0.418	0.741	0.421	0.734	0.440	0.705	0.436	0.689
		LEV	0.326	0.638	0.372	0.692	0.316	0.583	0.364	0.640
		BF	0.256	0.562	0.295	0.620	0.204	0.443	0.256	0.530
		FK	0.272	0.584	0.269	0.568	0.262	0.525	0.250	0.505
60	1/5 (10,50)	BRT	0.280	0.523	0.287	0.541	0.341	0.550	0.311	0.509
		LEV	0.198	0.390	0.274	0.520	0.202	0.361	0.278	0.488
		BF	0.144	0.310	0.209	0.436	0.105	0.227	0.188	0.380
		FK	0.186	0.385	0.169	0.355	0.187	0.362	0.161	0.313
	1/9 (6,54)	BRT	0.179	0.330	0.204	0.380	0.239	0.387	0.225	0.371
		LEV	0.115	0.212	0.213	0.389	0.120	0.206	0.220	0.374
		BF	<b>0.079</b>	<b>0.152</b>	0.153	0.313	<b>0.051</b>	<b>0.096</b>	0.142	0.274
		FK	0.124	0.234	<b>0.116</b>	<b>0.230</b>	0.133	0.235	<b>0.118</b>	<b>0.213</b>
120	1/1 (60,60)	BRT	0.709	0.963	0.711	0.965	0.676	0.931	0.667	0.926
		LEV	0.643	0.948	0.647	0.948	0.594	0.908	0.584	0.903



Table 3. Continued.

N	RoGS ( $n_1, n_2$ )	VR HT	Slightly-skewed- gamma distribution				Highly-skewed- gamma distribution			
			1:2	1:3	2:1	3:1	1:2	1:3	2:1	3:1
240	1/2 (40, 80)	BF	0.586	0.935	0.595	0.936	0.492	0.866	0.481	0.861
		FK	0.580	0.923	0.588	0.927	0.531	0.878	0.519	0.876
		BRT	0.667	0.948	0.664	0.947	0.654	0.915	0.624	0.898
		LEV	0.579	0.917	0.613	0.927	0.534	0.873	0.556	0.880
		BF	0.521	0.897	0.555	0.910	0.420	0.816	0.458	0.834
		FK	0.534	0.900	0.524	0.886	0.488	0.858	0.471	0.825
		BRT	0.482	0.819	0.489	0.818	0.508	0.785	0.477	0.753
		LEV	0.368	0.726	0.458	0.793	0.350	0.657	0.427	0.731
		BF	0.316	0.681	0.390	0.750	0.233	0.534	0.327	0.656
		FK	0.358	0.726	0.340	0.685	0.332	0.657	0.305	0.607
		BRT	0.339	0.636	0.348	0.645	0.402	0.639	0.357	0.589
		LEV	0.239	0.494	0.338	0.626	0.234	0.451	0.322	0.574
	1/5 (20, 100)	BF	0.192	0.425	0.274	0.565	0.137	0.303	0.235	0.488
		FK	0.246	0.507	0.228	0.479	0.235	0.470	0.198	0.416
		BRT	<b>0.929</b>	<b>0.999</b>	<b>0.923</b>	<b>0.999</b>	<b>0.882</b>	<b>0.996</b>	<b>0.881</b>	<b>0.995</b>
		LEV	0.906	0.998	0.899	<b>0.999</b>	0.852	<b>0.996</b>	0.851	<b>0.995</b>
	1/9 (12, 108)	BF	0.889	0.998	0.881	<b>0.999</b>	0.802	0.994	0.798	0.993
		FK	0.878	0.997	0.875	0.998	0.824	<b>0.996</b>	0.823	0.994
		BRT	0.901	0.999	0.892	<b>0.999</b>	0.856	0.993	0.845	0.991
		LEV	0.865	0.997	0.870	0.997	0.804	0.991	0.814	0.989
	1/1 (120, 120)	BF	0.843	0.997	0.849	0.996	0.736	0.985	0.757	0.986
		FK	0.846	0.997	0.834	0.995	0.784	0.990	0.767	0.985
		BRT	0.752	0.978	0.747	0.974	0.729	0.956	0.687	0.941
		LEV	0.660	0.961	0.716	0.966	0.604	0.927	0.650	0.932
	1/2 (80, 160)	BF	0.617	0.956	0.667	0.959	0.495	0.887	0.570	0.912
		FK	0.645	0.961	0.627	0.943	0.601	0.931	0.561	0.899
		BRT	0.574	0.902	0.583	0.890	0.591	0.866	0.538	0.828
		LEV	0.460	0.840	0.552	0.875	0.428	0.772	0.505	0.816
	1/5 (40, 200)	BF	0.410	0.813	0.494	0.851	0.310	0.673	0.415	0.769
		FK	0.459	0.848	0.448	0.807	0.430	0.789	0.387	0.732
		BRT								
		LEV								
	1/9 (24, 216)	BF								
		FK								
		BRT								
		LEV								

N: Total sample size; RoGS: Ratio of group size; HT: Homogeneity tests; VR: Variance ratio; BRT: Bartlett's test; LEV: Levene's test; BF: Brown-Forsythe test; FK: Fligner-Killeen test

Results in Table 3 indicated that under the gamma distribution, Bartlett's test tended to provide the best power among the homogeneity tests under all crossed conditions, followed by Levene's test. However, the Brown-Forsythe and Fligner-Killeen tests tended to yield the lowest power compared to the other homogeneity of variance tests considered in this study.

For the gamma distribution characterized by a slightly skewed shape, the maximum power values were generally attained under the condition of balanced group sizes 1/1 (120, 120), and in one case under the group size ratio of 1/2 (80, 160). Bartlett's test provided the highest power for the variance conditions of 1:2, 1:3, and 2:1 with the group size ratio of 1/1 (120, 120). For the variance condition of 3:1, Bartlett's, Levene's, and Brown-Forsythe tests all produced the highest power, with a value of 0.999, again for the balanced group size of 1/1 (120, 120). For the group size ratio of 1/2 (80, 160), Bartlett's test also yielded the highest power, with a value of 0.999 as in the previous cases.

For the gamma distribution characterized by a highly skewed shape, the maximum power values were attained under the balanced group sizes of 1/1 (120, 120), as in the slightly-skewed gamma distribution. Specifically, Bartlett's test yielded the highest power for the variance conditions of 1:2 and 2:1. For the 1:3 variance condition, Bartlett's, Levene's, and Fligner-Killeen tests yielded the same value of 0.996, while the Brown-Forsythe test produced a value of 0.994. For the variance condition of 3:1, Bartlett's and Levene's tests both yielded the same value of 0.995.

In both the slightly-skewed and highly-skewed distributions, the lowest power was observed under the group size ratio of 1/9 (6, 54). An analysis of the group variance ratios was conducted to ascertain the impact of the different methods under this condition. It was observed that the Brown-Forsythe test produced the lowest power in the 1:2 and 1:3 variance ratio conditions. In contrast, for the 3:1 and 2:1 variance ratio conditions, the Fligner-Killeen test yielded the lowest power values. Overall, the Brown-Forsythe and Fligner-Killeen tests provided the lowest power values when the sample size was minimal and the group size disparity was maximal (1/9, 6 vs. 54). These outcomes were consistent under both the slightly-skewed and highly-skewed gamma distributions. Furthermore, the power values under each variance condition demonstrated that the slightly-skewed gamma distribution yielded higher power than the highly-skewed distribution.

A comparative analysis of the methods revealed that Bartlett's test produced optimal results under all crossed conditions when the sample sizes were 60 and 120. Levene's test aligned closely with Bartlett's test. For the sample size of 240, Bartlett's test generally produced the highest values; however, the differences between methods were small, and there was a substantial increase in true rejection rates overall. Notably, in the highly-skewed distribution with group sizes 1/1 (120, 120) under the 1:3 variance condition, Bartlett's and Levene's tests provided the same values. In the slightly-skewed distribution under the 3:1 variance condition, Bartlett's, Levene's, and Brown-Forsythe tests all produced identical results, while the Fligner-Killeen test produced an almost identical value of 0.998.

As the difference between group sizes widened, the differences among methods also became more pronounced. When analyzed in terms of the lowest power under each crossed condition, the Brown-Forsythe and Fligner-Killeen tests generally produced the lowest values. Finally, as the total sample size increased, correct rejection rates also increased for all methods, resulting in higher power values. Altering the order of the variance ratios (1:2 versus 2:1; 1:3 versus 3:1) had only a minor impact on the results.

A general evaluation of the results regarding power values indicates that the correct rejection rates increased as the sample size increased when the group size remained constant. However, an increase in group size ratios led to a decrease in power values when the total sample size remained constant. It was established that an increase in the discrepancy between group variance ratios corresponded to an increase in the correct rejection rate. By contrast, no systematic change was observed when the order of group variance ratios was altered (1:2 versus 2:1; 1:3 versus 3:1).

The analysis revealed that Bartlett's test tended to yield the highest power values, while the Brown-Forsythe and Fligner-Killeen tests generally produced the lowest values. When the sample size was large (240) and the group size ratio was balanced (1/1), all tests provided close to, or even near-perfect, values. In contrast, when the sample size was modest (60) and the group size ratio was unbalanced (e.g., 1/9), a decline in power values was observed across all methods, with the disparities between methods becoming more pronounced. In such cases, Bartlett's test stood out for its consistently high power.

#### 4. DISCUSSION and CONCLUSION

This study compared the performance of Bartlett, Levene, Brown-Forsythe, and Fligner-Killeen tests in terms of error rates and power for two groups when the variances were equal (1:1) and

unequal (1:2, 1:3, 2:1, 3:1), across different distribution types (normal, skewed-normal, gamma), and varying sample sizes (60, 120, and 240) with different sample ratios (1/1, 1/2, 1/4, 1/9). In the first stage, the false rejection rates of the homogeneity tests were examined under normal, skewed-normal, and gamma distributions. When the variances for the two groups were equal under the normal distribution, the Fligner-Killeen test had the lowest false rejection rate when group sizes were balanced. In unbalanced cases, the Brown-Forsythe and Fligner-Killeen tests alternately produced the lowest false rejection rates. In other words, the type I error rates for Brown-Forsythe and Fligner-Killeen were lower than those of the other tests under the normal distribution. Similarly, Yi et al. (2020) stated that the Brown-Forsythe test was adequate for most population distribution shapes.

It should be noted that under the normal distribution, type I error rates across all simulation conditions were approximately 0.05 for all tests. However, under the skewed-normal and gamma distributions, the type I error rates for Bartlett's test and Levene's test increased more than those for the Brown-Forsythe and Fligner-Killeen tests. In particular, Bartlett's test showed the largest type I error rates under non-normal distributions. In other words, Bartlett's test was extremely sensitive to non-normal distributions and only performed well under the normal distribution. In parallel with this result, Chang et al. (2017) suggested using Bartlett's test only when the normality assumption is nearly certain. As in Yonar's (2024) research, Levene's test showed lower performance and larger type I error rates for normal distributions. In general, in some cases Levene's produced higher error rates than Bartlett's, while in other cases Bartlett's produced higher error rates than Levene's. However, a closer examination revealed that the results of the two tests were highly similar. A comparison of the slightly-skewed and highly-skewed distributions under both the skewed-normal and gamma distributions revealed that the highly-skewed distributions resulted in higher error rates. In general, greater skewness led to higher error levels as the sample size increased for all homogeneity of variance tests.

In the next stage, the results of the correct rejection rates of the homogeneity tests under normal, skewed-normal, and gamma distributions were examined. When the variances for the two groups were different under the normal distribution, Bartlett's test was generally the most powerful, followed by Levene's test. As Gastwirth et al. (2009) pointed out, Levene's test is a powerful method for situations where the normal distribution is not satisfied. The power of the Brown-Forsythe and Fligner-Killeen tests generally tended to decrease as the total sample size became smaller. Across all sample size conditions, the correct rejection rates of all tests were closer to each other when the group sizes were balanced. As the total sample size increased, the correct rejection rates increased for all methods. In parallel with the study of Wang et al. (2017), as the sample size increased, the correct rejection rates of all tests increased overall. Although the progression of the correct rejection rates of the methods varied under different sample conditions, it was evident that increasing the sample size had a positive effect on the performance of all methods.

For the normal, skewed-normal, and gamma distributions, the highest power values were observed when the sample size was large and the group sizes were balanced. Conversely, the lowest power values were observed when the sample size was smallest and the disparity between group sizes was greatest.

Depending on the conclusions of the current study, it is recommended that researchers select the most appropriate homogeneity of variance test based on the distribution, total sample size, group sample sizes, and variance ratios of the data. For example, if the sample size is 120, the data are normally distributed with a group variance ratio of 1:3, and the group size ratio is 1/5 (20, 100), Bartlett's test is a good choice when the focus is on power. The following recommendations are made:

- In general, the homogeneity of variance tests achieve perfect or near-perfect power as the total sample size increases and the disparity between group sizes decreases. Therefore, it is recommended that researchers use balanced and large samples whenever feasible.
- If type I error is a major concern in research, the Brown-Forsythe test should be preferred, particularly for non-normal distributions (skewed-normal and gamma). If the Brown-Forsythe test is not applicable, the Fligner-Killeen test may be considered a viable alternative.
- The Fligner-Killeen test is recommended for detecting small differences between group variances under a skewed-normal distribution.
- If there is an assumption or suspicion of heterogeneity of variance between groups, Bartlett's test may be preferred. If Bartlett's test is not applicable, Levene's test generally yields comparable results and may be considered a viable alternative without substantially reducing statistical power.
- If the group sizes are unbalanced and the sample size is small, Bartlett's test is recommended, as it tends to provide higher power values than the other methods.

As with all simulation studies, a limitation is that the results are only applicable to the conditions investigated in this research. To broaden the applicability of the current study, future researchers should conduct comparative analyses using real data.

### Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

### Contribution of Authors

**Serpil Çelikten Demirel:** Investigation, Resources, Visualization, and Writing-original draft. **Ayşenur Erdemir:** Methodology, Software, Formal Analysis, and Writing-original draft. **Esra Oyar:** Methodology, Formal Analysis, Validation, and Writing-original draft. **Tuba Gündüz:** Validation, and Writing-original draft.

### Orcid

Serpil Çelikten Demirel  <https://orcid.org/0000-0003-3868-3807>

Ayşenur Erdemir  <https://orcid.org/0000-0001-9656-0878>

Esra Oyar  <https://orcid.org/0000-0002-4337-7815>

Tuba Gündüz  <https://orcid.org/0000-0002-0921-9290>

### REFERENCES

- Abdullah N.F., & Muda, N. (2022). An overview of homogeneity of variance tests on various conditions based on type 1 error rate and power of a test. *Journal of Quality Measurement and Analysis*, 18(3), 111-130.
- Ahsanullah, M. (2017). *Characterizations of univariate continuous distributions* (Vol. 1). Amsterdam: Atlantis Press.
- Arsham, H., & Lovric, M. (2011). Bartlett's Test. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 87–88). Springer. [https://doi.org/10.1007/978-3-642-04898-2\\_132](https://doi.org/10.1007/978-3-642-04898-2_132)
- Arnold, B.C., Gómez, H.W., & Salinas, H.S. (2014). A doubly skewed normal distribution. *Statistics*, 49(4), 842-858. <https://doi.org/10.1080/02331888.2014.918618>
- Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12(2), 171-178. <http://www.jstor.org/stable/4615982>
- Bartlett, M.S. (1939, April). A note on tests of significance in multivariate analysis. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 35, No. 2, pp. 180-185). Cambridge University Press.



- Bono, R., Blanca, M.J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology*, 8, 1602. <https://doi.org/10.3389/fpsyg.2017.01602>
- Brown, M.B., & Forsythe, A.B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69(346), 364-367. <https://doi.org/10.1080/01621459.1974.104829557>
- Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance. I: Effect of inequality of variance in one-way classification. *Annals of Mathematical Statistics*, 25, 290–302. <http://www.jstor.org/stable/2236731>
- Chang, C.H., Pal, N., & Lin, J.J. (2017). A revisit to test the equality of variances of several populations. *Communications in Statistics-Simulation and Computation*, 46(8), 6360-6384. <https://doi.org/10.1080/03610918.2016.1202277>
- Conover, W.J., Johnson, M.E., & Johnson, M.M. (1981). A comparative study of test for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, 23, 351–361. <https://doi.org/10.2307/1268225>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.
- Fligner, M.A., & Killeen, T.J. (1976). Distribution-free two-sample tests for scale. *Journal of the American Statistical Association* 71(353), 210-213. <https://doi.org/10.1080/01621459.1976.10481517>
- Fraenkel, J., & Wallen, N. (2009). *How to Design and Evaluate Research in Education* (7th ed.) McGraw-Hill Education.
- Gamst, G., Meyers, L.S., & Guarino, A.J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS*. Cambridge University. <https://doi.org/10.1017/CBO9780511801648>
- Gastwirth, J.L., Gel, Y.R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, 24(3), 343-360. <http://dx.doi.org/10.1214/09-STS301>
- Glass, G.V. (1966). Testing homogeneity of variances. *American Educational Research Journal*, 3(3), 187-190. <https://doi.org/10.3102/00028312003003187>
- Gökpinar, E. (2022). Standardized likelihood ratio test for homogeneity of variance of several normal populations. *Communications in Statistics-Simulation and Computation*, 51(11), 6309-6319. <https://doi.org/10.1080/03610918.2020.1800037>
- Howell, D.C. (2010). *Statistical methods for psychology*. PWS-Kent Publishing Co.
- Katsileros, A., Antonetsis, N., Mouzaidis, P., Tani, E., Bebeli, P.J., & Karagrigoriou, A. (2024). A comparison of tests for homoscedasticity using simulation and empirical data. *Communications for Statistical Applications and Methods*, 31(1), 1-35. <https://doi.org/10.29220/CSAM.2024.31.1.001>
- Keppel, G., & Wickens, T.D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, Pearson Prentice Hall.
- Keskin, S. (2002). *Varyansların homojenliğini test etmede kullanılan bazı yöntemlerin I. tip hata ve testin gücü bakımından irdelenmesi* [An examination of some methods used in testing the homogeneity of variances in terms of type I error and test power] [Unpublished doctoral dissertation]. Ankara University.
- Kim, Y.J., & Cribbie, R.A. (2018). ANOVA and the variance homogeneity assumption: Exploring a better gatekeeper. *British Journal of Mathematical and Statistical Psychology*, 71(1), 1-12. <https://doi.org/10.1111/bmsp.12103>
- Kirk, R.E. (2008). *Statistics an Introduction* (5th ed.). Thomson Wadsworth
- Maitra, S.D. (1990). Skewness and the beta distribution. *Journal of the Operational Research Society*, 41(10), 953-961. <https://doi.org/10.1057/jors.1990.147>
- Orcan, F. (2020). Parametric or non-parametric: skewness to test normality for mean comparison. *International Journal of Assessment Tools in Education*, 7(2), 255-265. <https://doi.org/10.21449/ijate.656077>



- Öztürk, N.N. (2020). *Varyansların homojenliği için bazı testler ve karşılaştırmaları* [Some tests for the homogeneity of variances and comparisons] [Unpublished master dissertation]. Gazi University.
- Park, H.I. (2018). Tests of equality of several variances with the likelihood ratio principle. *Communications for Statistical Applications & Methods*, 25(4), 329-339. <https://doi.org/10.29220/CSAM.2018.25.4.329>
- Roscoe, J.T. (1975). *Fundamental Research Statistics for the Behavioural Sciences* (2nd ed.). Holt Rinehart & Winston.
- Sedgwick, P. (2015). A comparison of parametric and non-parametric statistical tests. *BMJ*, 350. <https://doi.org/10.1136/bmj.h2053>
- Shavelson, R.J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Pearson Education.
- Sarisoy, E.E., Potas, N., & Kara, M. (2013). A simulation study goodness-of-fit tests for the skewed normal distribution. In *Chaos, Complexity and Leadership 2012* (pp. 277-283). Springer Netherlands.
- Sarisoy, E.E., Potas, N., Kara, M. (2014). A simulation study goodness-of-fit tests for the skewed normal distribution. In S. Banerjee and Ş. Erçetin (eds). *Chaos, Complexity and Leadership 2012* (pp. 277-283). Springer Proceedings in Complexity. Springer, Dordrecht. [https://doi.org/10.1007/978-94-007-7362-2\\_36](https://doi.org/10.1007/978-94-007-7362-2_36)
- Tabachnick, B.G., & Fidell, L.S. (2007). *Using Multivariate Statistics* (5th ed.). Pearson.
- Wang, Y., Rodríguez de Gil, P., Chen, Y.H., Kromrey, J.D., Kim, E.S., Pham, T., Nguyen, D., & Romano, J.L. (2017). Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models. *Educ. Psychol. Meas.*, 77, 305–329. <https://doi.org/10.1177/0013164416645162>
- Woodbury, G. (2002). *An Introduction to Statistics: Improving Your Grade*. Belmont, CA: Brooks/Cole.
- Yi, Z., Chen, Y.H., Yin, Y., Cheng, K., Wang, Y., Nguyen, D., ... Kim, E. (2020). Brief Research Report: A Comparison of Robust Tests for Homogeneity of Variance in Factorial ANOVA. *The Journal of Experimental Education*, 90(2), 505-520. <https://doi.org/10.1080/00220973.2020.1789833>
- Yonar, A., Yonar, H., Demirsöz, M., & Tekindal, M.A. (2024). A comparative analysis for homogeneity of variance tests. *Journal of Science and Arts*, 24(2), 305-328. <https://doi.org/10.46939/J.Sci.Arts-24.2-a06>
- Zhou, Y., Zhu, Y., & Wong, K.Y. (2023). Statistical tests for homogeneity of variance for clinical trials and recommendations. *Contemporary Clinical Trials Communications*, 33, 101119. <https://doi.org/10.1016/j.conctc.2023.101119>