

e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by **DergiPark**

International Journal of Assessment Tools in Education

Volume: 11

Issue: 3

September 2024

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

Editor Dr. Omer KUTLU
Address Ankara University, Faculty of Educational Sciences, Cebeci Yerleşkesi,
Çankaya, Ankara, Türkiye

E-mail ijate.editor@gmail.com
omerkutlu@ankara.edu.tr

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
ijate.editor@gmail.com

Journal Contact Dr. Eren Can AYBEK
Address Department of Educational Sciences, Pamukkale University, Faculty of
Education, Kinikli Yerleşkesi, Denizli, 20070, Türkiye

Phone +90 258 296 31050
Fax +90 258 296 1200
E-mail erencanaybek@gmail.com

Address Dr. Anil KANDEMİR
Department of Educational Sciences, Agri Ibrahim Cecen University,
Faculty of Education, Agri, Türkiye
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 author(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)
- Education Resources Information Center (ERIC)
- TR Index (ULAKBIM),
- EBSCOhost,
- SOBIAD,
- JournalTOCs,
- MIAR (Information Matrix for Analysis of the Journals),
- idealonline,
- CrossRef,

Editor

Dr. Omer KUTLU, *Ankara University, Türkiye*

Section Editors

Dr. Erkan Hasan ATALMIS, *Manisa Celal Bayar Üniversitesi, Türkiye*
Dr. Ebru BALTA, *Ağrı İbrahim Çeçen Üniversitesi, Türkiye*
Dr. Safiye BILICAN DEMİR, *Kocaeli University, Türkiye*
Dr. Selma SENEL, *Balikesir University, Türkiye*
Dr. Esin YILMAZ KOGAR, *Nigde Omer Halisdemir University, Türkiye*
Dr. Fatma Betül KURNAZ, *Karabük Üniversitesi, Türkiye*
Dr. Sumeyra SOYSAL, *Necmettin Erbakan University, Türkiye*

Editorial Board

Dr. Beyza AKSU DUNYA, *Bartın University, Türkiye*
Dr. Stanislav AVSEC, *University of Ljubljana, Slovenia*
Dr. Kelly D. BRADLEY, *University of Kentucky, United States*
Dr. Okan BULUT, *University of Alberta, Canada*
Dr. Javier Fombona CADAVIECO, *University of Oviedo, Spain*
Dr. Seockhoon Chung, *University of Ulsan, Korea*
Dr. R. Nukhet CIKRIKCI, *İstanbul Aydın University, Türkiye*
Dr. William W. COBERN, *Western Michigan University, United States*
Dr. Nuri DOGAN, *Hacettepe University, Türkiye*
Dr. Selahattin GELBAL, *Hacettepe University, Türkiye*
Dr. Anne Corinne HUGGINS-MANLEY, *University of Florida, United States*
Dr. Francisco Andres JIMENEZ, *Shadow Health, Inc., United States*
Dr. Nicole KAMINSKI-OZTURK, *The University of Illinois at Chicago, United States*
Dr. Tugba KARADAVUT, *Izmir Democracy University, Türkiye*
Dr. Orhan KARAMUSTAFAOGLU, *Amasya University, Türkiye*
Dr. Yasemin KAYA, *Atatürk 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. Froilan D. MOBO, *Ama University, Philippines*
Dr. Hamzeh MORADI, *Sun Yat-sen University, China*
Dr. Nesrin OZTURK, *Izmir Democracy University, Türkiye*
Dr. Turan PAKER, *Pamukkale University, Türkiye*
Dr. Murat Dogan SAHIN, *Anadolu University, Türkiye*
Dr. Hossein SALARIAN, *University of Tehran, Iran*
Dr. Halil İbrahim SARI, *Texas A&M University, United States*
Dr. Ragip TERZI, *Harran University, Türkiye*
Dr. Turgut TURKDOGAN, *Pamukkale University, Türkiye*
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. Ayse BILICIOGLU GUNES, *TED 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

Harmonizing perspectives to understand attitudes: A mixed methods approach to crafting an assessment literacy attitude scale

Page: 424-444 [PDF](#)

Beyza Aksu, Stefanie Wind, Mehmet Can Demir

An investigation into the effect of different missing data imputation methods on IRT-based differential item functioning

Page: 445-462 [PDF](#)

Fatma Ünal, Hakan Koğar

Adaptation of the quiet quitting scale for teachers to Turkish culture: An empirical psychometric investigation

Page: 463-480 [PDF](#)

Müslim Alanoğlu, Songül Karabatak, Alper Uslukaya, Ayşenur Kuloğlu

The validity and reliability study of the theory of mind inventory-2 (TOMI-2) Turkish version

Page: 481-506 [PDF](#)

Canan Keleş Ertürk, Kezban Tepeli

The effect of rater training on rating behaviors in peer assessment among secondary school students

Page: 507-523 [PDF](#)

Nazira Tursynbayeva, Umur Öç, İsmail Karakaya

How many grades of response categories does the commitment to the profession of medicine scale provide the most information?

Page: 524-536 [PDF](#)

Murat Tekin, Çetin Toraman, Ayşen Melek Aytuğ Koşan

Educating non-specialized audiences about seismic design principles using videos and physical models

Page: 537-566 [PDF](#)

Mauricio Morales-beltran, Ecenur Kızılörenli, Ceren Duyal

Algebraic knowledge for teaching test: An adaptation study

Page: 567-588 [PDF](#)

Ali Bozkurt, Begüm Özmusul

The mental imagery scale for art students: Building and validating a short form

Page: 589-607 [PDF](#)

Handan Narin Kızıltan, Hatice Cigdem Bulut

The use of ChatGPT in assessment

Page: 608-621 [PDF](#)

Mehmet Kanik

Harmonizing perspectives to understand attitudes: A mixed methods approach to crafting an assessment literacy attitude scale

Beyza Aksu Dünya^{1,2}, Stefanie A. Wind², Mehmet Can Demir^{3*}

¹Bartın University, Faculty of Education, Department of Educational Sciences, Bartın, Türkiye

²University of Alabama, College of Education, Tuscaloosa, AL, USA

³Bartın University, Faculty of Education, Department of Educational Sciences, Bartın, Türkiye

ARTICLE HISTORY

Received: Feb. 20, 2024

Accepted: Apr. 29, 2024

Keywords:

Assessment literacy,
Higher education,
Scale development,
Mokken scale analysis,
Mixed-methods.

Abstract: Assessment literacy's vital role in faculty effectiveness within higher education lacks sufficient tools for measuring faculty attitudes on this matter. Employing a sequential mixed-methods approach, this study utilized the theory of planned behavior to develop the Assessment Literacy Attitude Scale (ALAS) and evaluate its psychometric properties within the U.S. higher education context. The qualitative phase involved a literature review of relevant studies and existing self-report measures, interviews with stakeholders, and panel reviews to shape initial item development. Following the establishment of a conceptual foundation and a comprehensive overview of the scale's construction, our study advanced to the quantitative stage that involves factor analytical and item response theory approaches using data from 260 faculty across three public universities in the U.S. Exploratory factor analysis (EFA) was employed initially to obtain preliminary insights into the scale's factorial structure and dimensionality. Confirmatory factor analysis (CFA) was subsequently applied with separate data and the findings largely supported the conclusions from the EFA. Exploratory and confirmatory factor analyses resulted in 15 items loading across two factors in a good model fit range. Finally, we used nonparametric item response theory (IRT) techniques based on Mokken Scale Analysis (MSA) to evaluate individual items for evidence of effective psychometric properties to support the interpretation of ALAS scores, including monotonicity, scalability, and invariant item ordering. The newly-developed scale shows promise in assessing faculty attitudes toward enhancing their assessment literacy.

1. INTRODUCTION

In the realm of higher education, faculty members consistently face challenges related to student assessment (Jankowski & Marshall, 2017; Medland, 2019; Sadler, 2017). The complexities associated with assessment practices are further intensified by the current demand for accountability in higher education (Caspersen & Smeby, 2018; Liu et al., 2012; Scholl & Olsen, 2014). The shift toward outcome-based education (Adam, 2004; Coates, 2016; Singh & Ramya, 2011) alongside changes in quality assurance and accreditation standards (Williams, 2016) has spurred a renewed emphasis on assessment literacy (Biggs & Tang, 2011; Dann, 2014;

*CONTACT: Mehmet Can Demir ✉ mehmetdemir@bartin.edu.tr 📍 Bartın University, Faculty of Education, Department of Educational Sciences, Bartın, Türkiye.

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

e-ISSN: 2148-7456

Eubanks, 2019; Wolf et al., 2015). The rapid advancements in artificial intelligence (AI) and its influence on assessment also necessitate faculty members to adapt and enhance their assessment methods to embrace these changes (McMurtrie, 2023). Despite this emphasis, assessment literacy remains an area in need of improvement in higher education (Pastore, 2022).

As a practical response to the growing and evolving importance of assessment literacy among faculty, higher education institutions have made recent strides in providing various faculty development opportunities to enhance assessment literacy among faculty. One of the objectives of such programs is to encourage changes in faculty members' routine assessment practices (Hines, 2009; Holmboe et al., 2011). To foster faculty members' willingness to adopt these practices, it is also important to explore their attitudes towards enhancing their assessment literacy. Understanding faculty attitudes towards their own assessment literacy and motivation to improve in this area can inform the development and improvement of in professional development initiatives. For example, institutions can tailor workshops, seminars, or training programs to address specific areas for improvement, leading to continuous growth in assessment practices. Tailored professional development opportunities play a vital role in ensuring that faculty members involved in assessment are well-prepared to carry out their responsibilities proficiently (Horst & Prendergast, 2020).

Accrediting bodies often require higher education institutions to provide evidence of effective assessment practices (Dill, 2007). Having a well-established scale to measure faculty attitudes and demonstrate changes in their attitudes towards assessment literacy can provide evidence that an institution is committed to fostering a culture of continuous improvement in assessment. By using a psychometrically sound scale to measure faculty members' attitudes towards enhancing their assessment literacies, higher education institutions can effectively make data-driven decisions in designing targeted professional development initiatives and assess their impact on faculty attitudes. Recognizing the link between attitudes and behavior (Ajzen, 1991), such a scale can also offer evidence of institutions' dedication to fostering a culture of continuous improvement in assessment practices. Understanding faculty attitudes towards enhancing assessment literacy allows institutions to tailor professional development programs to cater to individual strengths and areas for improvement. Customized initiatives aligned with faculty attitudes can support improvement in engagement and motivation, resulting in improved retention of knowledge and skills, making the training more productive and impactful.

1.1. Assessment Literacy and Higher Education

Assessment literacy in higher education is a multifaceted construct with a comprehensive scope, encompassing various dimensions and components. In general terms, it is envisioned as an individual's thorough understanding of assessment requirements (Zhu & Evans, 2022). This multidimensional construct includes a spectrum of skills, knowledge, and dispositions (Pastore & Andrade, 2019). Initially, assessment literacy was narrowly defined to establish a common language regarding assessment terminology. However, the conceptual boundaries have expanded significantly over time. For example, Price et al. (2012) broadened the definition by incorporating principles, methods, standards, and feedback. Subsequent contributions by Xu and Brown (2016) introduced the identification of assessors as a crucial component, emphasizing knowledge, conceptions, and practice. Evans (2016) further enriched the concept by incorporating an affective domain, highlighting the inclusion of staff and student entitlement in assessment literacy. Furthermore, the cognitive dimension of assessment literacy, as detailed by Balloo et al. (2018), underscores the significance of making assessment criteria explicit and transparent, thereby clarifying the requirements of assessment tasks. In essence, assessment literacy in higher education encompasses a diverse range of dimensions, mirroring its rich and evolving nature. Pastore and Andrade's (2019) three-dimensional model further elucidates

assessment literacy, highlighting conceptual knowledge, the practical application of this knowledge to support learning, and a socio-emotional dimension.

Recent studies, such as those by Kremmel and Harding (2020), have extended the scope of assessment literacy to encompass socio-cultural values, personal beliefs, and attitudes. This evolution demonstrates a transition from a foundational understanding rooted in terminology and knowledge domains to a more intricate and holistic concept that integrates socio-cultural and personal dimensions. Within higher education, assessment literacy entails managing assessment practices and upholding standards and fairness (Zhu & Evans, 2022).

Significantly, the acknowledgment of personal attitudes and beliefs emerges as a crucial element within the assessment literacy framework. Assessment literacy is being influenced by personal beliefs, attitudes, and conceptions on assessment (O'Neill et al., 2023). Cultivating assessment literacy and promoting faculty ownership of this evolving definition necessitates the recognition and understanding of the nuanced individual attitudes and beliefs held by educators. By doing so, initiatives aimed at enhancing assessment literacy can be precisely tailored to align with the diverse perspectives and values that faculty bring to the educational setting. This approach fosters a more inclusive and effective stance toward assessment practices. In this context, the Theory of Planned Behavior (TPB) (Ajzen, 1991), with its well-grounded principles, holds substantial potential for guiding these efforts.

1.2. Purpose

The purpose of this study is to develop the Assessment Literacy Attitude Scale (ALAS) and evaluate its psychometric properties as a tool for measuring faculty attitudes towards assessment literacy enhancement. Specifically, we focused on the following research questions:

1. What is the internal structure of the initial set of ALAS items?
2. What modifications or refinements can be made to improve the psychometric properties of the ALAS based on results from the Exploratory Factor Analysis?
3. What is the degree of reproducibility of the ALAS items' structure?
4. What is the degree to which ALAS items conform to invariant item ordering principles?
 - 4.1. What is the degree of monotonicity exhibited by individual items within the scale?
 - 4.2. How scalable are the items within the scale, and what does this reveal about their ability to discriminate between different levels of the latent trait?
 - 4.3. Does the scale exhibit invariant item ordering, indicating consistent item difficulty across different levels of the latent trait?
5. Does ALAS yield sufficient reliability evidence?

1.3. Theoretical Framework

We grounded development of our scale items to TPB (Ajzen, 1991). The TPB stands as a highly influential framework for predicting and explaining human behavior, as outlined by Ajzen (1991, 2001). It has demonstrated successful applications in diverse domains, including professional development and adult and lifelong learning, where it proves valuable in comprehending the link between attitude and behavior (Archie et al., 2022; Dunn et al., 2018; Kao et al., 2018; Madigan & Kim, 2021). The central emphasis of the theory lies in an individual's intention to carry out a specific behavior, which, in our context, pertains to participating in activities and adopting programs and strategies to enhance assessment literacy. Intentions are regarded as the driving force behind behavior. Generally, a more robust intention to undertake a behavior correlates with a higher likelihood of successfully completing the action (Ajzen, 1991). The theory posits that the intention to adopt a behavior involves several psychological stages. These include developing a positive attitude towards the behavior, forming beliefs about the behavior's value, influenced by others' approval or disapproval, and engaging in the behavior based on perceived competency or the absence of constraints.

Attitude toward the behavior pertains to the extent of positive or negative evaluation of the behavior under consideration. This implies that a more nuanced and specific attitude serves as a more accurate predictor for the targeted outcome behavior in question (Ajzen & Timko, 1986). In the context of assessment literacy, discerning the extent to which faculty members appraise their active involvement in enhancing assessment literacy becomes crucial for forecasting their intent to participate in faculty development activities/initiatives related to assessment. Subjective norms encompass the perceived social standards that impact whether individuals sense external pressure to engage in a particular behavior. Multiple research studies affirm a positive correlation between perceived norms and behaviors among adults (Hora & Anderson, 2012; Knauder & Koschmieder, 2019; Rimal & Real, 2003). Finally, perceived behavioral control pertains to an individual's assessment of the ease or difficulty associated with performing a particular behavior, considering any constraints that may exist. The more challenging individuals perceive it to be to initiate or complete the behavior, the less likely they are to develop strong intentions to do so. Key issues related to perceived behavioral control in the context of enhancing assessment literacy would involve exploring whether faculty members believe that developing assessment literacy is within their sphere of influence/competency.

2. METHOD

2.1. Instrument Development

In accordance with the principles outlined by Creswell and Plano Clark (2011) for the development of an exploratory instrument, the initial step in our instrument development procedure involved establishing a clear definition of the construct. This process aimed to identify the main themes, constructs, and available instruments related to assessment literacy in the context of higher education. Drawing insights from existing literature, particularly in the context of assessment literacy within higher education, we started by understanding how researchers defined this construct in previous studies. Additionally, we engaged in semi-structured face-to-face interviews with two individuals affiliated with the faculty development office at the U.S. university where the research was initiated. Through these interviews, we sought to gain perspectives on the attitudes of faculty members towards assessment literacy, further enriching our exploration. The scale's initial development primarily centered on the dimensions of the theory of planned behavior framework, as it succinctly explains the motivational influences on behavior (Conner & Armitage, 1998). To establish an initial set of items aligned with the adopted theory, we drew upon the themes identified through an analysis of both the relevant literature and the insights gained from interviews. During this item generation process, we focused on ensuring sufficient content coverage. Each item underwent scrutiny for language and content appropriateness, considering clarity, length, and relevance to the target population. Furthermore, we assessed each item for potential biases, leading or suggestive phrasing, loading (encouraging automatic answers), and double-barreled content. These efforts resulted in a final set of 29 items. Our hypothesis posited that each of the 29 items would fall within one of the three domains of the TPB.

We presented the 29-item scale to two experts in assessment and measurement, who are currently working as faculty in different public universities in the U.S. They examined each item for relevance, accuracy, and representativeness, using a three-category rating scale (1= should be deleted, 2= requires revision, and 3= can be used) to affirm content and face validity. We also sought their suggestions on the number of response categories. We incorporated input from these two experts to complete the refinement of the instrument before its administration to the developmental sample for a think-aloud session (discussed in the next section). Following the panel reviews that identified redundancy as the primary concern, we subsequently downsized the initial instrument from 29 items to a more streamlined version containing 20 items.

To investigate examinee response processes for the new items, we conducted concurrent think-aloud sessions using the pilot instrument consisting of 20 items. In these sessions, three participants engaged in real-time verbalization of their thoughts and reactions while responding to the items via Zoom (2023). The think-aloud sessions followed the principles outlined by Padilla and Leighton (2017), where participants were requested to articulate their thoughts without interruption or leading questions from the interviewer. This psychological method, aligned with the Standards (AERA et al., 2014), is designed to capture data on human information processing and responses. The primary objective was to gain insights into participants' cognitive processes, allowing us to refine the instrument based on their feedback. Consequently, adjustments were made to the wording of some items to enhance clarity, informed by the observations and feedback obtained during these sessions.

2.2. Participants

The target population for this study was faculty members who have experience teaching or are currently teaching in either graduate or undergraduate classes in the U.S. setting. To reach a representative sample from the target population, we employed a combination of two non-probabilistic sampling techniques: convenience sampling and snowball sampling (Cochran, 1977). The only demographic information we collected was academic discipline, as this characteristic was important for the focus of our study. By limiting the demographic variables to academic discipline, we aimed to prioritize the relevance and specificity of the findings to the academic and teaching contexts under investigation.

The study involved a total of 260 faculty members from three U.S. public universities. In the initial round of data collection during the summer of 2023, participants were recruited from faculty at two large public universities in the southern region. For our initial analysis using Exploratory Factor Analysis (EFA), a total of 136 individuals responded to the instrument out of the 1687 faculty invited (8.06% completion rate). Despite the substantial data collection efforts and response rates, 33 individuals did not complete the instrument due to various reasons. Participants included 29 faculty from Natural and Applied Sciences (28.2%), 33 from Social Sciences (32%), 27 from Humanities (26.2%), and 14 from Business (13.6%). A subsequent round of data collection, utilizing the same sampling approach, was conducted for the Confirmatory Factor Analysis (CFA) and phases. The second round of data collection took place in the fall of 2023, involving faculty from a different public university in the Midwest. Over 2000 faculty were invited. The second sample included 197 respondents, with 40 failing to complete the instrument. Faculty representation was as follows: 51 from Natural and Applied Sciences (32.5%), 37 from Social Sciences (23.5%), 43 from Humanities (27.4%), 21 from Business (13.4%), and 5 from other (or multiple) academic disciplines (3.2%).

2.3. Data Collection Procedures

The study was conducted in the United States, where higher education institutions actively prioritize the continuous development of faculty member programs to enhance assessment literacy. Upon obtaining Institutional Review Board (IRB) approval, face-to-face interviews were conducted with faculty development professionals on campus, delving into their perspectives. Subsequently, the think-aloud process described earlier was implemented via Zoom with three participants, who offered valuable qualitative insights. Following these qualitative phases, the scale administration employed a web-based recruitment strategy through Qualtrics (Dillman et al., 2014). Faculty members were invited to participate via a weblink, which was disseminated across various channels, including college faculty listservs, ResearchGate, and LinkedIn. Moreover, participants were encouraged to share the survey link within their professional networks through social media, creating a snowball sampling effect that fostered broader participation.

The scale, structured into three sections, began with a comprehensive informed consent presentation in the initial section. Subsequently, participants engaged with the main set of scale

items. The final segment prompted participants to indicate their discipline grouping (e.g., humanities, applied sciences, etc.). Participants responded to scale items using an ordinal four-category Likert-type scale ranging from strongly disagree to strongly agree ($1 = \text{strongly disagree}$, $2 = \text{disagree}$, $3 = \text{agree}$, $4 = \text{strongly agree}$).

Throughout the data collection process, participants were assured of the confidentiality and exclusive research-oriented use of their responses and scale data. To uphold participant privacy, the assessment procedure did not request any personal identifiers, such as names or other identifiable information.

2.4. Data Analysis

For the instrument development stage of the study, the data analysis approach included content analysis of existing literature. This method ensured the provision of content evidence of validity for the instrument. Following expert reviews, a standardized statistical approach was applied to gather evidence related to content validity based on expert review. Specifically, we calculated inter-rater agreement statistics to evaluate the reliability of the expert review process.

Both factor analytic and non-parametric Item Response Theory (IRT) based approaches were employed to gather evidence related to the internal structure of the ALAS items. We provide details about our analysis related to each approach in the following sections.

2.4.1. Phase 1: Exploratory factor analysis

Prior to executing EFA, the assumptions of sampling adequacy (Kaiser-Meyer-Olkin [KMO] test) for evaluating sample size sufficiency and Bartlett's Test of Sphericity to ensure adequate item correlation were examined. EFA was performed using responses from 103 participants to the 20-item version of the ALAS (Table 1). We performed the analysis using the "psych" package for R (R Core Team, 2023; Revelle, 2023).

Table 1. 20-item version of ALAS.

-
- I1. I continually strive to enhance my assessment literacy.
 - I2. I must stay current with the latest assessment methods to fulfill my teaching responsibilities.
 - I3. I believe that improving my assessment literacy is crucial to enhance student learning outcomes.
 - I4. I feel motivated to learn more about assessment strategies to better teach my students.
 - I5. I am open to exploring new assessment techniques to improve my teaching practices.
 - I6. I believe that having strong assessment literacy is important for being an effective faculty member.
 - I7. I believe that increasing my assessment literacy will help me to better meet the needs of a diverse student population.
 - I8. I view increasing my assessment literacy as a continuous process, rather than a one-time task.
 - I9. Frequent conversation with colleagues improves my assessment practices.
 - I10. I value learning new concepts about assessment.
 - I11. Faculty professional development in assessment is necessary for quality instruction.
 - I12. I would like to complete more training in assessment in the future.
 - I13. I would only take an assessment training if it was required by my department.
 - I14. I plan to continue learning new techniques about assessment.
 - I15. I participate in professional development activities regarding assessment.
 - I16. I seek out opportunities to increase my assessment literacy.
 - I17. Learning innovative assessment approaches is valuable.
 - I18. I strive to use different applications and technology in assessment.
 - I19. Learning new tools and information in assessment is part of my professional development.
 - I20. I think faculty in higher education should have substantial knowledge in assessment.
-

During the EFA, we utilized specific criteria for retaining items. These criteria included: (a) a measure of sampling adequacy (KMO) of 0.5 or higher for each item, according to Field (2000), (b) a statistically significant Bartlett's Test of Sphericity value (Pett et al., 2003) and (c) adhering to Howard's (2016) recommendation that each item should have a minimum loading of 0.40 onto its primary factor, a maximum loading of 0.30 onto other factors, and a minimum difference of 0.20 loading between the primary factor and other factors.

The overall KMO value was equal to 0.914 and Bartlett Test of Sphericity value was statistically significant—indicating that the item responses could be explored using EFA ($\chi^2 = 1314.921$, $df=190$, $p<0.001$). For most items, the standardized univariate skewness and kurtosis values fell outside the range of ± 1.96 . The multivariate normality was checked by Mardia's test for multivariate normality and the multivariate skewness and kurtosis values were statistically significant ($p<0.001$), which indicates univariate and multivariate non-normal response distributions. As such, we applied EFA using the minimum residuals method suggested by Kline (1994).

2.4.2. Phase 2: Confirmatory factor analysis

To avoid overfitting in scale development studies, it is recommended to conduct CFA on a separate sample to confirm the structure of the proposed scale that resulted from an EFA (Fokkema & Greiff, 2017). Therefore, CFA was performed using responses from a new sample of 157 participants. The analysis was performed using the “lavaan” package for R (Rosseel, 2012).

Upon examination of the univariate and multivariate normality values, it was found that most of the standardized univariate skewness and kurtosis values fall out range of ± 1.96 range. The multivariate normality was checked by Mardia's test for multivariate normality and the multivariate skewness and kurtosis values were statistically significant ($p<0.001$). Consequently, we applied the diagonally weighted least squares (DWLS) (Muthén, 1993) estimation method for the CFA since it is a suitable estimator with small samples and non-normal distributions.

Numerous fit indices are utilized in the CFA domain, with the Comparative Fix Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) being the most prevalent. The recommended criteria for best and good fit according to Hu and Bentler (1999) are as follows: CFI and TLI should be equal to or greater than 0.95 for best fit and equal to or greater than 0.90 for good fit; RMSEA should be less than or equal to 0.05 for best fit and less than or equal to 0.08 for good fit; and SRMR should be equal to or greater than 0.05 for best fit and equal to or greater than 0.10 for good fit. The internal consistency of the scale was evaluated by computing Cronbach's alpha (α ; Cronbach, 1951) and McDonald's omega (ω ; McDonald, 1999).

2.4.3. Phase 3: Nonparametric item response theory: Mokken scale analysis

Our first step in exploring the ALAS under nonparametric IRT framework was to examine the items for evidence of psychometric quality using basic item analysis statistics. First, we examined the frequency of responses in each rating scale category across items to ensure that we could apply our selected item analysis techniques to the data. Then, we examined item responses for evidence of internal consistency using inter-item and corrected item-total correlations. We conducted these analyses within the factors identified in the earlier analysis phases.

Next, we evaluated the scaling properties of the ALAS items within the identified factors using MSA (Mokken, 1971), which is a theory-driven nonparametric approach to item response theory (IRT). We used MSA to evaluate the ALAS items for several reasons. First, MSA includes several graphical and statistical techniques that provide an exploratory perspective into the degree to which items exhibit fundamental scaling properties while maintaining an ordinal

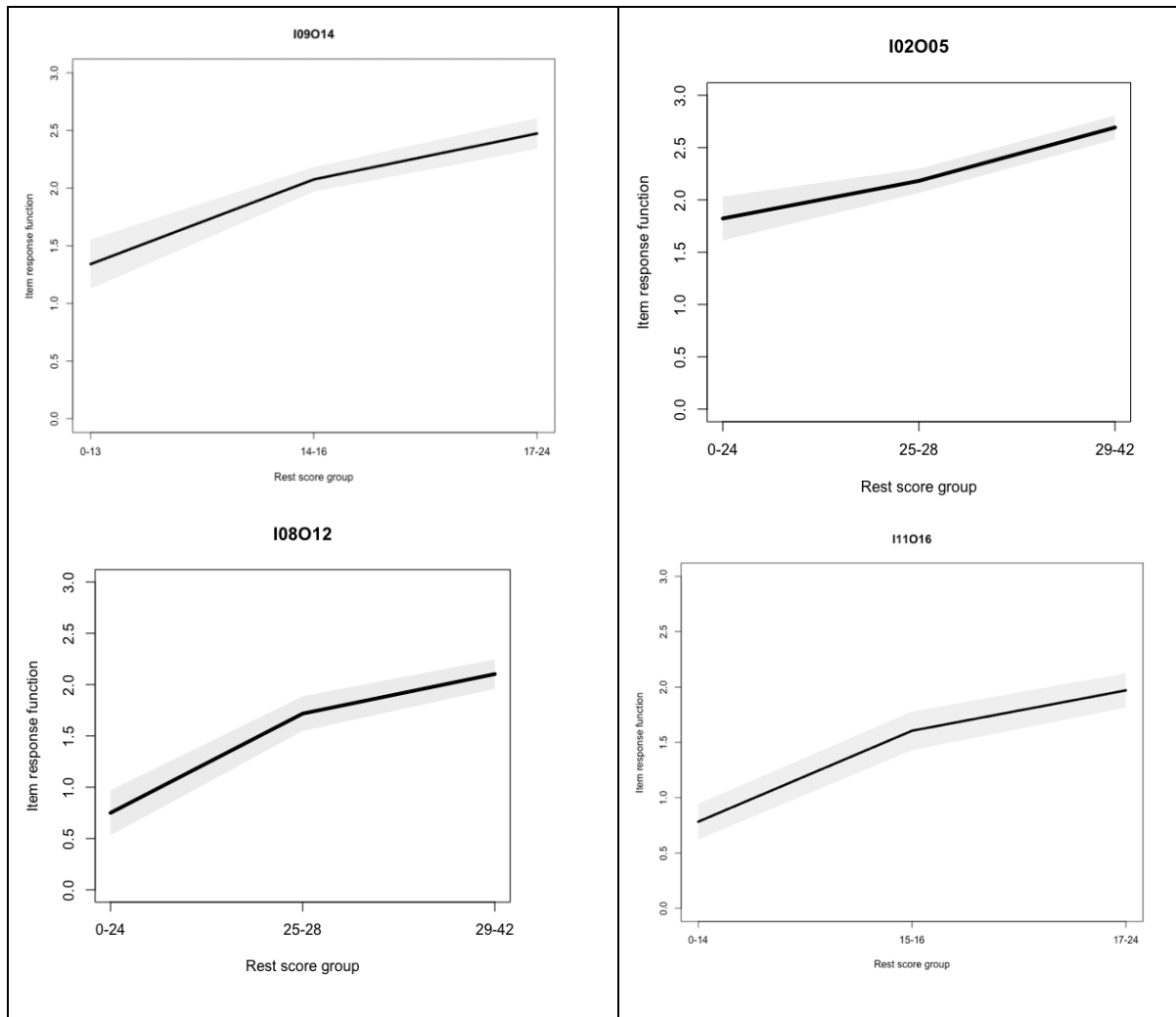
level of measurement. This is particularly useful in scale development studies for which the construct or a set of items is not well-understood, and the use of an interval-level scale may be inappropriate or unnecessary (Meijer et al., 2015). Although it is nonparametric, MSA is characterized by clear ordering requirements for the relationship between item and person characteristics (i.e., the item response function; IRF) that reflect invariant measurement. In contrast to a-theoretical nonparametric IRT techniques (e.g., kernel smoothing) (Mazza et al., 2014; Ramsay & Silverman, 2005), these requirements provide a framework in which to evaluate item quality. Finally, because MSA is nonparametric, it can be used with relatively small examinee sample sizes, such as the current sample.

Typical MSA procedures are based on two nonparametric scaling models: (1) the Monotone Homogeneity Model (MHM), and (2) the Double Monotonicity Model (DMM). These models are characterized by ordering requirements that facilitate item analysis. The MHM is based on three requirements: (1) unidimensionality: a single latent variable is sufficient to explain most of the variation in item responses, (2) local item independence: After controlling for the primary latent variable, there are no meaningful associations between item responses (i.e., responses are statistically independent), and (3) monotonicity: participants' average responses for individual items are non-decreasing as their locations on the latent variable increase. The DMM shares the same requirements as the MHM and adds a fourth requirement: invariant item ordering (IIO): items have the same relative difficulty order for all participants. We used techniques based on these models to examine three major indicators of measurement quality for the ALAS items. From the MHM, we examined evidence of item monotonicity and item scalability. From the DMM, we examined evidence of invariant item ordering (IIO). We conducted the MSA analyses using the “mokken” package for R (van der Ark, 2007, 2012). Details on the specifics and procedures for testing these requirements are outlined below.

2.4.3.1. Item Monotonicity. For individual items, monotonicity occurs when participants' average ratings on an item are non-decreasing as their locations on the latent variable increase. Unlike parametric IRT models for which participant locations on the latent variable are estimated using an interval scale, MSA uses an ordinal nonparametric indicator of person locations based on total scores. Specifically, item monotonicity is evaluated using item-specific restscores, which are total scores minus participant scores on the item of interest. Typical procedures for evaluating item monotonicity include combining participants with equal or adjacent restscores into restscore groups with approximately balanced sample sizes in each group to improve statistical power for evaluating item properties.

Figure 1 illustrates item monotonicity at the overall item level using nonparametric IRFs for two example items from the ALAS. In each plot, the x-axis shows examinee rest-score groups, and the y-axis shows the rating scale, which was re-scaled to start at zero. The nonparametric IRFs show the average ratings for each item within restscore groups, and light shading around the line shows a 95% confidence interval. Both items exhibited adequate monotonicity because the average ratings are non-decreasing as rest-scores increase. In addition to graphical displays, researchers can also evaluate monotonicity using one-sided statistical hypothesis tests that evaluate whether monotonicity holds between pairs of adjacent restscore groups.

We examined monotonicity for each item within each of the the ALAS factors using graphical displays of nonparametric IRFs similar to Figure 1 as well as statistical hypothesis tests.

Figure 1. Examples of plots for evaluating item monotonicity.

2.4.3.2. Item Scalability. In the context of MSA, scalability refers to the degree to which response patterns associated with individual or groups of items support a consistent interpretation of item ordering across persons. Specifically, scalability coefficients describe the degree to which item responses are free from Guttman errors, or unexpected response patterns given item and person ordering on the latent variable. MSA procedures include scalability coefficients for individual items (H_i), pairs of items (H_{ij}), and sets of three or more items (H). Researchers typically interpret scalability coefficients as an indicator of overall item quality and fit to the MHM (Sijtsma & Molenaar, 2002). In scale development studies, researchers often focus on scalability coefficients for individual items, which can be calculated as:

$$H_i = 1 - \frac{\sum_{j \neq i} F_{ij}}{\sum_{j \neq i} E_{ij}} \quad (1)$$

where F_{ij} is the observed frequency of Guttman errors associated with item i in combination with all other items in the scale, and E_{ij} is the expected frequency of Guttman errors for item i based on marginal independence. Researchers typically interpret item scalability coefficients with values greater than or equal to $H_i = 0.30$ as evidence of meaningful contribution to a scale (Meijer & Baneke, 2004; Sijtsma & van der Ark, 2017).

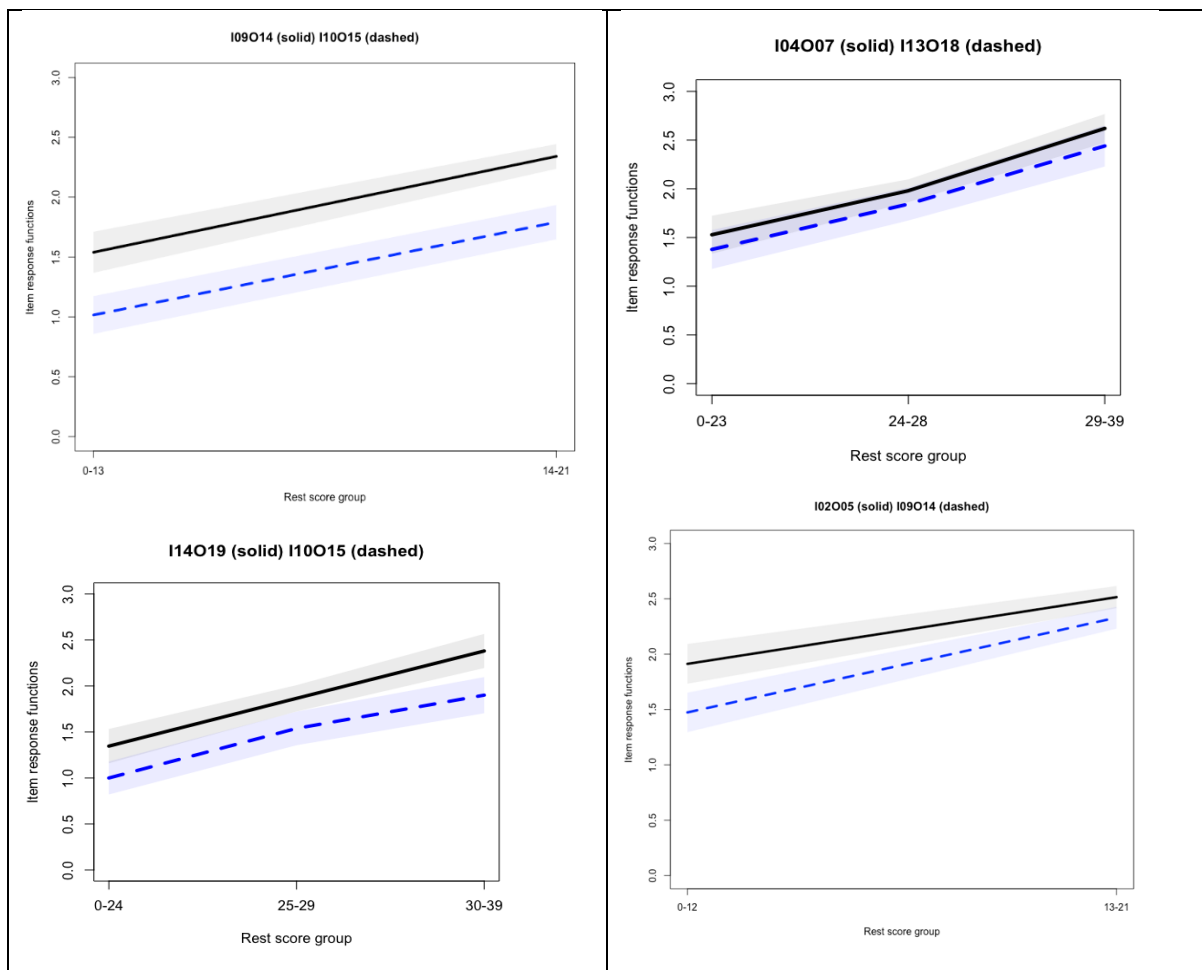
2.4.3.3. Invariant Item Ordering. The last major category of MSA item analysis is Invariant Item Ordering (IIO). This property is related to the DMM, and it describes the degree to which items exhibit a consistent relative difficulty ordering for participants with different locations on the latent variable. IIO has important theoretical and practical implications that are

relevant for scale development. When items adhere to IIO, there is a single and consistent hierarchy of items that does not depend on examinees' location on the latent variable. In practice, researchers typically evaluate IIO for rating scale items such as those included in the ALAS using Manifest IIO analyses (Ligtvoet et al., 2010), which involve evaluating pairs of nonparametric IRFs for evidence of non-intersection across examinee restscores specific to the item pair.

For example, Figure 2 shows plots for evaluating IIO using items from the ALAS. For these IIO analyses, restscores are calculated specific to the item pair of interest, using examinee total scores across all items minus their scores on the two items being evaluated. In each plot, separate IRFs are plotted for each item that show examinees' average response to each item within rest-score groups calculated using their total score on all of the ALAS items except IO9014 and I10015. These two items adhere to IIO because the IRF for IO9014 (solid line) is always higher—indicating higher average ratings—compared to the IRF for I10015 (dashed line). Adherence to IIO for these two items suggests that participants always endorse IO9014 more readily than I10015. IIO also holds for the item pair made up of IO2005 and IO9014.

Figure 1 illustrates item monotonicity at the overall item level using nonparametric IRFs for two example items from the ALAS. In each plot, the x-axis shows examinee rest-score groups, and the y-axis shows the rating scale, which was re-scaled to start at zero. The nonparametric IRFs show the average ratings for each item within restscore groups, and light shading around the line shows a 95% confidence interval. Both items exhibited adequate monotonicity because the average ratings are non-decreasing as rest-scores increase. In addition to graphical displays, researchers can also evaluate monotonicity using one-sided statistical hypothesis tests that evaluate whether monotonicity holds between pairs of adjacent restscore groups.

Figure 2. Examples of plots for evaluating invariant item ordering.



3. FINDINGS

3.1. Initial Scale Construction Results

In our examination of existing measures and studies on assessment literacy in higher education, the literature review yielded a pool of 40 candidate items. Through extensive discussions and careful consideration, some items were excluded from the pool for various reasons, such as redundancy. Following this refinement process, we initiated expert reviews, commencing with a set of 29 candidate items. The main themes identified from the content review, encompassing perceived necessity and the rationale for enhancing assessment literacy, were identified in conjunction with insights from face-to-face interviews with two experts. Additionally, identified several instruments related to skill improvement, such as the Effective Lifelong Learning Inventory (ELLI) (Crick et al., 2004), but noted that these were not specifically designed to measure assessment literacy.

After collecting expert item-level ratings, we computed Cohen's Kappa (Cohen, 1960) as a chance-corrected measure of inter-rater agreement for each criterion. According to Landis and Koch's (1977) guidelines, we achieved an almost perfect agreement level, surpassing 0.80 across the criteria of relevance, accuracy, and representativeness. This result indicates a high degree of consensus. Notably, during this process, nine items were identified and subsequently removed from consideration. This decision was rooted in the rating of 1 given by two raters on at least one of the criteria, ensuring a stringent and consistent approach to item selection.

3.2. EFA Results

The overall KMO value was equal to 0.914 and Bartlett Test of Sphericity value was statistically significant—indicating that the item responses could be explored using EFA ($\chi^2 = 1314.921$, $df = 190$, $p < 0.001$). For most items, the standardized univariate skewness and kurtosis values fell outside the range of ± 1.96 (see Table 2). and the multivariate skewness and kurtosis values were statistically significant ($p < 0.001$), which indicates univariate and multivariate non-normal response distributions. As such, we applied EFA using the minimum residuals method suggested by Kline (1994) with oblique rotation which allows for correlation between the latent factors.

Table 2. Descriptive statistics for EFA.

Item	Mean	Mdn	SD	Skewness	Kurtosis	Item	Mean	Mdn	SD	Skewness	Kurtosis
I1	2.954	3	0.802	-0.581	0.128	I11	3.075	3	0.832	-0.642	-0.109
I2	2.926	3	0.782	-0.467	0.007	I12	2.861	3	0.901	-0.423	-0.557
I3	3.159	3	0.791	-0.759	0.274	I13	2.269	2	0.953	0.358	-0.746
I4	2.925	3	0.843	-0.529	-0.175	I14	3.074	3	0.68	-0.82	1.753
I5	3.333	3	0.684	-1.073	1.954	I15	2.636	3	0.84	-0.098	-0.546
I6	3.056	3	0.818	-0.729	0.263	I16	2.704	3	0.788	-0.351	-0.152
I7	3.206	3	0.774	-0.873	0.653	I17	3.167	3	0.634	-1.047	3.366
I8	3.299	3	0.69	-0.999	1.716	I18	2.981	3	0.785	-0.439	-0.167
I9	2.843	3	0.888	-0.337	-0.622	I19	2.907	3	0.746	-0.54	0.391
I10	3.167	3	0.69	-0.752	1.211	I20	3.259	3	0.661	-0.733	1.147

The EFA results supported a two-factor structure, unlike the grounded TPB with 3 factors. Also, the Velicer's minimum average partial (MAP) (Velicer et al., 2000) test and parallel analysis supported two-factor structure (Figure 3). The two factors jointly captured 52.5% of the variance in the set of items and were positively correlated with each other ($r = 0.66$). Based on Howard's (2016) rule, five items (#1, #2, #4, #9, and #13) were removed (see Table 3). After removal of the items, the EFA was re-run and the factor structure and loadings were similar. Cronbach α values were equal to 0.91 and 0.88 for factor #1 and factor #2, respectively—

suggesting strong internal consistency. For discriminant validity, average variance extracted (AVE) values for factors were acceptable (0.49 and 0.60) and composite reliability (CR) values or factors were good (0.89 and 0.90) (Fornell & David, 1981). Also, the heterotrait-monotrait ratio of correlations (HTMT) was lower than 0.85 threshold (0.82) and this indicates the structure has sufficient discriminant validity (Henseler et al., 2015). Following the interpretation of these results and necessary revisions, a distinct sample was employed for CFA.

Figure 3. Scree plot.

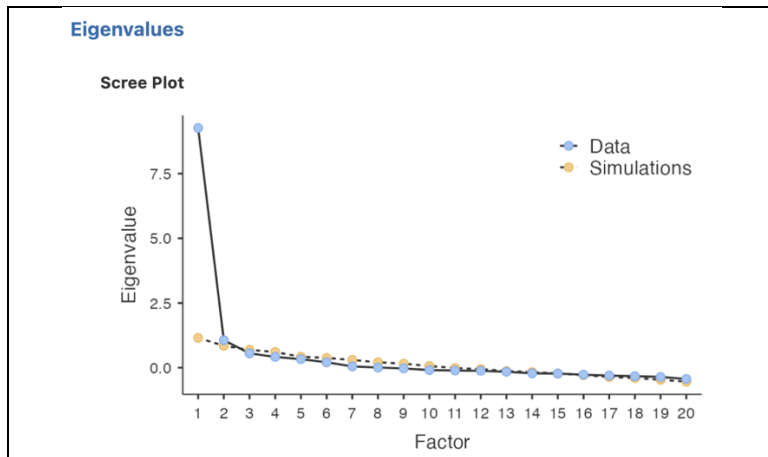


Table 3. Factor loadings for EFA.

	Item	Factor #1	Factor #2
I19	Learning new tools and information in assessment is part of my professional development.	0.837	
I14	I plan to continue learning new techniques about assessment.	0.832	
I5	I am open to exploring new assessment techniques to improve my teaching practices.	0.801	
I8	I view increasing my assessment literacy as a continuous process, rather than a one-time task.	0.774	
I18	I strive to use different applications and technology in assessment.	0.794	-0.138
I10	I value learning new concepts about assessment.	0.610	0.185
I16	I seek out opportunities to increase my assessment literacy.	0.569	0.260
I17	Learning innovative assessment approaches is valuable.	0.542	0.234
I15	I participate in professional development activities regarding assessment.	0.410	
<i>I1</i>	<i>ally strive to enhance my assessment literacy.</i>	<i>0.366</i>	<i>0.347</i>
I11	Faculty professional development in assessment is necessary for quality instruction.	-0.170	0.846
I3	I believe that improving my assessment literacy is crucial to enhance student learning outcomes.		0.840
I6	I believe that having strong assessment literacy is important for being an effective faculty member.		0.822
I12	I would like to complete more training in assessment in the future.	0.124	0.638
I20	I think faculty in higher education should have substantial knowledge in assessment.	0.134	0.624
I7	I believe that increasing my assessment literacy will help me to better meet the needs of a diverse student population.	0.206	0.621
<i>I4</i>	<i>I feel motivated to learn more about assessment strategies to better teach my students.</i>	<i>0.425</i>	<i>0.486</i>
<i>I2</i>	<i>I must stay current with the latest assessment methods to fulfill my teaching responsibilities.</i>	<i>0.314</i>	<i>0.468</i>
<i>I9</i>	<i>Frequent conversation with colleagues improves my assessment practices.</i>	<i>0.259</i>	<i>0.279</i>
<i>I13</i>	<i>I would only take an assessment training if it was required by my department.</i>		<i>-0.269</i>

Note: italicized items were removed

3.3. CFA Results

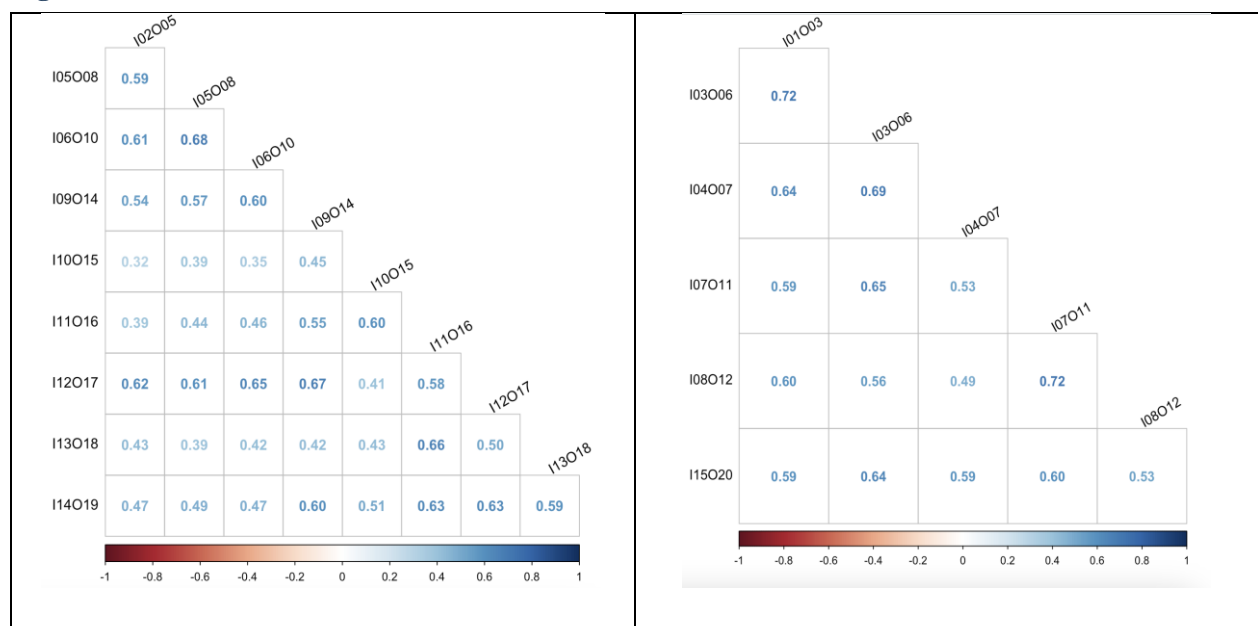
The descriptive statistics and reliability coefficients for the two dimensions of the scale are presented in Table 4. With the CFA sample, the Spearman correlation coefficients (ρ) between all pairs of items were within the range of $0.32 \leq \rho \leq 0.72$ and all of these values were statistically significant ($p < 0.001$). Matrix representations of the correlations are provided in Figure 4.

Table 4. Descriptive statistics and reliability coefficients.

Factor	Item*	Mean	Mdn	SD	Skewness	Kurtosis	α	ω	Mean (F)	SD (F)
#1	I02O05	3.287	3	0.651	-0.789	1.387	0.917	0.921	2.944	0.569
	I05O08	3.28	3	0.696	-0.789	0.719				
	I06O10	3.141	3	0.74	-0.715	0.581				
	I09O14	3	3	0.716	-0.743	1.096				
	I10O15	2.478	3	0.764	-0.098	-0.349				
	I11O16	2.51	3	0.773	-0.074	-0.355				
	I12O17	3.089	3	0.664	-0.766	1.709				
	I13O18	2.879	3	0.827	-0.321	-0.466				
	I14O19	2.854	3	0.749	-0.216	-0.284				
#2	I01O03	2.955	3	0.728	-0.537	0.458	0.916	0.917	2.939	0.646
	I03O06	3.019	3	0.791	-0.589	0.099				
	I04O07	3.032	3	0.729	-0.552	0.407				
	I07O11	2.904	3	0.791	-0.616	0.275				
	I08O12	2.647	3	0.833	-0.478	-0.268				
	I15O20	3.064	3	0.74	-0.679	0.65				

*: I02O05 stands for Item #2 (Old Item #5), I05O08 stands for Item #5 (Old Item #8), etc.

Figure 4. Inter-item correlation matrices.



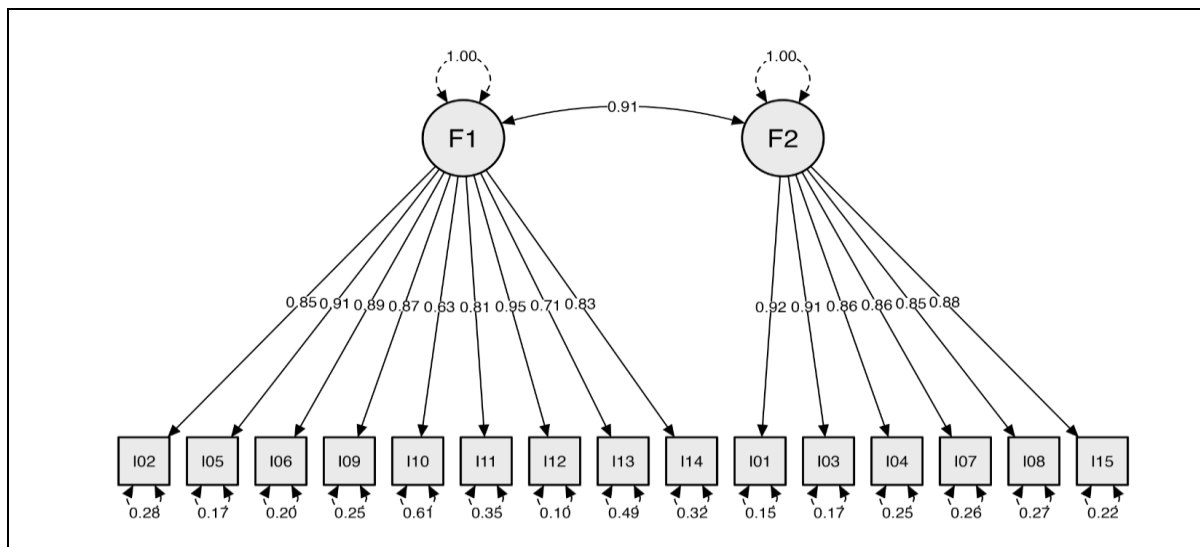
To be sure of the number of dimensions of the model, CFA was conducted with one-factor and two-factor models. The comparison was in favor of two-factor model and the results were given in Table 5.

Table 5. Comparison of one-factor and two-factor models.

Model	χ^2	df	CFI	RMSEA	SRMR	ΔX^2	Δdf	ΔCFI	$\Delta RMSEA$
One-factor	262.077	90	0.992	0.112	0.076	49.46***	1	0.003	0.017
Two-factor	212.617	89	0.995	0.095	0.070				

According to the CFA results, the model was statistically significant ($\chi^2= 212.617$, $df= 89$, $p<0.001$). While some of the fit indices were in the good fit range (CFI= 0.995, TLI= 0.994), others were in acceptable fit range (SRMR= 0.070) or mediocre (slightly below the acceptable fit range; RMSEA= 0.095) (Hu & Bentler, 1999). Factor loadings for individual items ranged from 0.63 to 0.95, the correlation between factors was equal to 0.91, and all of the values were statistically significant ($p<0.001$). Figure 5 illustrates these results using a path diagram.

Figure 5. Path diagram.



3.4. MSA Results

Preliminary data screening revealed that the inter-item and corrected item-total correlations were positive for all items within each factor. Moreover, internal consistency statistics indicated acceptable levels of internal consistency reliability (Factor 1: $\alpha= 0.92$; $\omega_{hierarchical}= 0.95$; Factor 2: $\alpha= 0.92$; $\omega_{hierarchical}$, total =0.94). Together, these results support further analysis of the ALAS items using a scaling approach.

Within each factor, all of the ALAS items adhered to monotonicity and IIO with no statistically significant violations. As shown in Table 6, all items had positive scalability coefficients. For Factor 1, individual item scalability coefficients ranged from $H_i = 0.55$ ($SE = 0.07$) for item I10O15 to $H_i = 0.72$ ($SE = 0.04$) for item I09O14. The overall scalability coefficient for the ALAS items in Factor 1 was equal to $H = 0.67$ ($SE = 0.04$). For Factor 2, individual item scalability coefficients ranged from $H_i = 0.68$ ($SE = 0.05$) for item I04O07 to $H_i = 0.75$ ($SE = 0.04$) for item I08O12. The overall scalability coefficient for Factor 2 was equal to $H = 0.73$ ($SE = 0.04$).

Table 6. Item scalability coefficients.

Factor #1			Factor #2		
Item	Item Scalability (H_i)	Standard Error	Item	Item Scalability (H_i)	Standard Error
I02O05	0.67	0.05	I01O03	0.74	0.04
I05O08	0.66	0.05	I03O06	0.75	0.04
I06O10	0.67	0.05	I04O07	0.69	0.05
I09O14	0.72	0.04	I07O11	0.73	0.04
I10O15	0.55	0.07	I08O12	0.75	0.04
I11O16	0.70	0.04	I15O20	0.70	0.05
I12O17	0.74	0.04			
I13O18	0.61	0.06			
I14O19	0.71	0.04			

3.5. Summary of the Findings

In reconciling the outcomes of factor analysis and MSA, we identified and confirmed two factors, deviating from the initially hypothesized three within the TPB framework. Nevertheless, it is crucial to note that these two factors align with the core constructs of the theory. The explanation for each factor is provided below.

Factor 1: Attitude in Learning (new approaches, tools, etc.). This factor appears to align closely with the "attitude" construct in the TPB. Participants' attitude in learning new approaches and tools likely encompasses their personal evaluations of the benefits and drawbacks associated with adopting new assessment approaches. This factor could influence faculty's inclination to embrace assessment literacy practices.

Factor 2: Perceived Importance of Assessment Literacy (AL). This factor can be seen as a combination of the subjective norms and perceived behavioral control constructs from the TPB. The perceived importance of assessment literacy may reflect social influences (subjective norms) where educators gauge the significance of assessment literacy based on external factors such as colleagues' opinions or institutional priorities. Additionally, the perceived importance of assessment literacy could encompass a sense of control over the behavior (perceived behavioral control), as faculty may believe that developing assessment literacy is a critical factor within their sphere of influence.

4. DISCUSSION

This study aims to develop a psychometrically-sound assessment literacy attitude scale for educators, especially in higher education sector. We addressed a series of questions to fulfill this aim. In addressing the first research question, the EFA results revealed a two-factor structure, deviating from the anticipated three factors posited by the TPB. This unexpected outcome underscores the complexity inherent in the domain of assessment literacy enhancement. Despite this departure from the anticipated structure, the findings suggest a reinterpretation of the TPB in our context. The elements within the TPB framework persist in significance, despite their reconfigured arrangement, with attitudes toward adopting new assessment approaches reflecting individual inclinations and the perceived importance of assessment literacy. This encapsulates both subjective norms and perceived behavioral control. The realignment underscores the complex nature of fostering assessment literacy within the TPB framework in the context of higher education.

Subsequent to the EFA, a careful inspection of factor loadings led to the removal of five items (Research question [RQ] #2). Two items, one of which was negatively-worded, were discarded

due to their failure to load onto any factor. The negative formulation, as repeatedly cautioned in the literature (Chang, 1995; Cole et al., 2019; Sliter & Zickar, 2014; Wright & Masters, 1982), raised concerns about the item's interpretability by respondents and fit of the data. Notably, cautionary evidence from a study by Sliter and Zickar (2014) employing IRT framework highlighted that the scales comprising only positively-worded items yielded more information at the peak of the information curve and across a wider range of the trait scale. Moreover, trait scores were estimated with smaller standard errors under such conditions. In addition, think-aloud session insights revealed discrepancies in the length of participant discussions for the other removed item, emphasizing the importance of incorporating qualitative methods in the scale development process (Morell & Tan, 2009; Zhou, 2019).

The proposed factor structure, derived from the EFA findings, underwent scrutiny via CFA with a distinct sample using the 15-item version of the ALAS instrument (RQ #3). The CFA results displayed commendable fit indices, characterized by favorable CFI and TLI values. However, a slight deviation was observed in the RMSEA, resting at 0.095. In conducting both the EFA and CFA, we adhered to the commonly recommended guideline of a minimum of 5 respondents per item (Tabachnick & Fidell, 1996). However, despite meeting this criterion, our RMSEA, a measure assessing how well the model reproduces the observed data, fell within the medium range. This outcome underscores the influence of sample size on fit indices, as larger sample sizes tend to yield more precise estimates, potentially leading to lower RMSEA values. This observation emphasizes the significance of considering sample size implications in interpreting CFA outcomes and points towards a prospective avenue for future research to explore the robustness of the factor structure across diverse sample sizes.

Within the domain of scale development studies, the integration of both classical test theory (CTT) and item response theory (IRT) stands as a crucial and frequently employed practice for a comprehensive assessment of psychometric properties (Dilek & Akbaş, 2022; McKown et al. 2023; Wright & Jenkins-Guarnieri, 2023). In alignment with this methodological paradigm, our study underscores the importance of using multiple measurement approaches to elucidate the underlying constructs of our assessment literacy instrument. The synergistic use of EFA, CFA, and MSA, was designed to establish a robust foundation for comprehending the factor structure and measurement properties. In pursuit of this objective, we extended our inquiry to MSA (RQ #4), further enriching the depth of our psychometric evaluation. All of the ALAS items demonstrated monotonicity and IIO without statistically significant violations, thereby enhancing the interpretability of our assessment literacy instrument.

5. CONCLUSION

Our study contributes a distinctive perspective on the application of Theory of Planned Behavior (TPB) constructs within the realm of advancing assessment literacy in higher education. The findings underscore the perceived significance of enhancing assessment literacy in facilitating the adoption of faculty development programs, innovative assessment methodologies, and tools. This perceived significance is propelled by various factors, including social influence stemming from institutional priorities and the recognition of assessment as a pivotal determinant of faculty influence. Higher education institutions can capitalize on faculty perceptions by strategically elevating the place of assessment among institutional priorities. This emphasis should be tangibly manifested through a multifaceted approach, including targeted assessment workshops in diverse formats, specialized training modules, and hands-on practice sessions. To enhance accessibility and support, institutions may consider incorporating artificial intelligence tools, such as chatbots, to provide prompt assistance in assessment-related queries. Such a technologically-driven support system, available 24/7, may empower faculty members with real-time guidance and resources, fostering a dynamic and responsive culture of assessment literacy.

While this study has yielded valuable insights, it is crucial to acknowledge its limitations. Firstly, the reliance on a relatively small sample size, comprising faculty members from public sector R1 universities, may restrict the generalizability of findings to broader populations within higher education, particularly in diverse contexts like private universities and teaching-based institutions. Future studies should encompass more varied university settings to ensure a comprehensive understanding. The limited scope of participants might not fully capture the diverse perspectives and experiences prevalent in larger academic environments. In the context of assessment practices, a gap in the literature still pertains to the relationship between faculty members' planned assessment enhancement behavior and their attitudes. This aspect requires further exploration and research to enrich the existing body of knowledge. Additionally, the use of self-report measures introduces a potential source of bias, as participants may respond based on perceived beliefs rather than providing objective assessments of their behavior. While the study makes significant contributions to the comprehension of assessment literacy and faculty development, these limitations underscore the necessity for future research with larger and more diverse samples. Incorporating objective measures, such as an assessment literacy level test, will further enhance the robustness of the findings.

Acknowledgments

This research was financially supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under grant number 1059B192201517.

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:** University of Alabama, 23-04-6561.

Contribution of Authors

Beyza Aksu Dünya: Funding, Literature review, Conception, Design, Investigation, Methodology, Item writing, Data collection, Receiving experts' opinions, and Writing-original draft. **Stefanie A. Wind:** Statistical analysis, Supervision, and Critical review. **Mehmet Can Demir:** Literature review, Conception, Methodology, Data interpretation, Statistical analysis, and Writing-original draft.

Orcid

Beyza Aksu Dünya  <https://orcid.org/0000-0003-4994-1429>

Stefanie A. Wind  <https://orcid.org/0000-0002-1599-375X>

Mehmet Can Demir  <https://orcid.org/0000-0001-7849-7078>

REFERENCES

- Adam, S. (2004). *Using learning outcomes: A consideration of the nature, role, application and implications for European education of employing "learning outcomes" at the local, national and international levels*. Paper presented at the Bologna Seminar, Heriot-Watt University, Edinburgh United Kingdom. http://www.aic.lv/ace/ace_disk/Bologna/Bol_s_emin/Edinburgh/S_Adam_Bacgrerep_presentation.pdf Accessed on 16 November 2023.
- 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)
- Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, 52(1), 27-58. <https://doi.org/10.1146/annurev.psych.52.1.27>
- Ajzen, I., & Timko, C. (1986). Correspondence between health attitudes and behavior. *Basic and Applied Social Psychology*, 7(4), 259-276. https://doi.org/10.1207/s15324834baspo704_2
- Archie, T., Hayward, C.N., Yoshinobu, S., & Laursen, S.L. (2022). Investigating the linkage between professional development and mathematics instructors' use of teaching practices

- using the theory of planned behavior. *Plos One*, 17(4), e0267097. <https://doi.org/10.1371/journal.pone.0267097>
- Baloo, K., Norman, M., & Winstone, N.E. (2018, January). Evaluation of a large-scale inclusive assessment intervention: a novel approach to quantifying perceptions about assessment literacy. In *The Changing Shape of Higher Education-Can Excellence and Inclusion Cohabit?: Conference Programme and Book of Abstracts*. University of Southern Queensland. https://srhe.ac.uk/arc/conference2018/downloads/SRHE_Conf_2018_Programme_Papers.pdf
- Biggs, J., & Tang, C. (2011). Train-the-trainers: Implementing outcomes-based teaching and learning in Malaysian higher education. *Malaysian Journal of Learning and Instruction*, 8, 1-19.
- Caspersen, J., & Smeby, J.C. (2018). The relationship among learning outcome measures used in higher education. *Quality in Higher Education*, 24(2), 117-135. <https://doi.org/10.1080/13538322.2018.1484411>
- Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement*, 55(6), 991-997. <https://doi.org/10.1177/0013164495055006007>
- Coates, H. (2016). Assessing student learning outcomes internationally: Insights and frontiers. *Assessment & Evaluation in Higher Education*, 41(5), 662-676. <https://doi.org/10.1080/02602938.2016.1160273>
- Cochran, W.G. (1977). *Sampling techniques*. John Wiley & Sons.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46. <https://doi.org/10.1177/001316446002000104>
- Cole, K.L., Turner, R.C., & Gitchel, W.D. (2019). A study of polytomous IRT methods and item wording directionality effects on perceived stress items. *Personality and Individual Differences*, 147(6), 63-72. <https://doi.org/10.1016/j.paid.2019.03.046>
- Conner, M., & Armitage, C.J. (1998). Extending the theory of planned behavior: A review and avenues for further research. *Journal of Applied Social Psychology*, 28(15), 1429-1464. <https://doi.org/10.1111/j.1559-1816.1998.tb01685.x>
- Creswell, J.W., & Clark, V.P. (2011). *Mixed methods research*. SAGE Publications.
- Crick, R.D., Broadfoot, P., & Claxton, G. (2004). Developing an effective lifelong learning inventory: The ELLI project. *Assessment in Education: Principles, Policy & Practice*, 11(3), 247-272. <https://doi.org/10.1080/0969594042000304582>
- Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. <https://doi.org/10.1007/BF02310555>
- Dann, R. (2014). Assessment as learning: blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, 21(2), 149-166. <https://doi.org/10.1080/0969594X.2014.898128>
- Dilek, H., & Akbaş, U. (2022). Investigation of education value perception scale's psychometric properties according to CTT and IRT. *International Journal of Assessment Tools in Education*, 9(3), 548-564. <https://doi.org/10.21449/ijate.986530>
- Dill, D. (2007). *Quality assurance in higher education: Practices and issues*. The 3rd International Encyclopedia of Education.
- Dunn, R., Hattie, J., & Bowles, T. (2018). Using the Theory of Planned Behavior to explore teachers' intentions to engage in ongoing teacher professional learning. *Studies in Educational Evaluation*, 59, 288-294. <https://doi.org/10.1016/j.stueduc.2018.10.001>
- Eubanks, D. (2019). Reassessing the elephant, part 1. *Assessment Update*, 31(2), 6-7. <https://doi.org/10.1002/au.30166>
- Evans, C. (2016). *Enhancing assessment feedback practice in higher education: The EAT framework*. University of Southampton. <https://www.southampton.ac.uk/assets/importe>

- [d/transforms/content-block/UsefulDownloads_Download/A0999D3AF2AF4C5AA24B5BEA08C61D8E/EAT%20Guide%20April%20FINAL1%20ALL.pdf](#)
- Field, A. (2003). *Discovering Statistics using IBM SPSS statistics*. Sage Publications.
- Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it [Editorial]. *European Journal of Psychological Assessment*, 33(6), 399–402. <https://doi.org/10.1027/1015-5759/a000460>
- Fornell, C., & David, F.L. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18, 39-50. <https://doi.org/10.2307/3151312>
- 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 Marketing Science*, 43, 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hines, S.R. (2009). Investigating faculty development program assessment practices: What's being done and how can it be improved?. *The Journal of Faculty Development*, 23(3), 5.
- Holmboe, E.S., Ward, D.S., Reznick, R.K., Katsufakis, P.J., Leslie, K.M., Patel, V.L., ... & Nelson, E.A. (2011). Faculty development in assessment: the missing link in competency-based medical education. *Academic Medicine*, 86(4), 460-467. <https://doi.org/10.1097/acm.0b013e31820cb2a7>
- Hora, M.T., & Anderson, C. (2012). Perceived norms for interactive teaching and their relationship to instructional decision-making: A mixed methods study. *Higher Education*, 64, 573-592. <https://doi.org/10.1007/s10734-012-9513-8>
- Howard, M.C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?. *International Journal of Human-Computer Interaction*, 32(1), 51-62. <https://doi.org/10.1080/10447318.2015.1087664>
- Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Jankowski, N.A., & Marshall, D.W. (2017). *Degrees that matter: Moving higher education to a learning systems paradigm*. Routledge. <https://doi.org/10.4324/9781003444015>
- Kao, C.P., Lin, K.Y., & Chien, H.M. (2018). Predicting teachers' behavioral intentions regarding web-based professional development by the theory of planned behavior. *EURASIA Journal of Mathematics, Science and Technology Education*, 14(5), 1887-1897. <https://doi.org/10.29333/ejmste/85425>
- Kline, P. (1994). *An easy guide to factor analysis*. Routledge.
- Knauder, H., & Koschmieder, C. (2019). Individualized student support in primary school teaching: A review of influencing factors using the Theory of Planned Behavior (TPB). *Teaching and Teacher Education*, 77, 66-76. <https://doi.org/10.1016/j.tate.2018.09.012>
- Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, 17(1), 100-120. <https://doi.org/10.1080/15434303.2019.1674855>
- 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>
- Ligtvoet, R., Van der Ark, L.A., Marvelde, J.M. te, & Sijtsma, K. (2010). Investigating an Invariant Item Ordering for Polytomously Scored Items. *Educational and Psychological Measurement*, 70(4), 578–595. <https://doi.org/10.1177/0013164409355697>
- Liu, O.L., Bridgeman, B., & Adler, R.M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, 41(9), 352-362. <https://doi.org/10.3102/0013189X12459679>

- Madigan, D.J., & Kim, L.E. (2021). Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit. *Teaching and Teacher Education*, 105, 103425. <https://doi.org/10.1016/j.tate.2021.103425>
- Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in Item Response Theory. *Journal of Statistical Software*, 58(6). <https://doi.org/10.18637/jss.v058.i06>
- McDonald, R.P. (1999). *Test theory: A unified treatment*. Taylor & Francis.
- McKown, C., Kharitonova, M., Russo-Ponsaran, N.M., & Aksu-Dunya, B. (2023). Development and Validation of a shortened form of SELweb EE, a Web-Based Assessment of Children's Social and Emotional Competence. *Assessment*, 30(1), 171-189. <https://doi.org/10.1177/107319112111046044>
- Medland, E. (2019). 'I'm an assessment illiterate': Towards a shared discourse of assessment literacy for external examiners. *Assessment & Evaluation in Higher Education*, 44(4), 565-580. <https://doi.org/10.1080/02602938.2018.1523363>
- Meijer, R.R., & Baneke, J.J. (2004). Analyzing psychopathology items: A case for Nonparametric Item Response Theory Modeling. *Psychological Methods*, 9(3), 354–368. <https://doi.org/10.1037/1082-989X.9.3.354>
- Meijer, R.R., Tendeiro, J.N., & Wanders, R.B.K. (2015). The use of nonparametric item response theory to explore data quality. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to typical performance assessment* (pp. 85–110). Routledge.
- Mokken, R.J. (1971). *A theory and procedure of scale analysis*. De Gruyter.
- Morell, L., & Tan, R.J.B. (2009). Validating for use and interpretation: A mixed methods contribution illustrated. *Journal of Mixed Methods Research*, 3(3), 242-264. <https://doi.org/10.1177/1558689809335079>
- Muthén, B.O. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen, & J.S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Sage Publishing.
- O'Neill, G., McEvoy, E., & Maguire, T. (2023). Supporting assessment literacy in changing times. In C. Evans and M. Waring (Eds.), *Research handbook on innovations in assessment and feedback in higher education*. Elgar Publishing.
- Padilla, J.L., & Leighton, J.P. (2017). Cognitive interviewing and think aloud methods. In B. Zumbo & A. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211-228). Springer.
- Pastore, S. (2022). Assessment Literacy in the higher education context: A critical review. *Intersection: A Journal at the Intersection of Assessment and Learning*, 4(1). <https://doi.org/10.61669/001c.39702>
- Pastore, S., & Andrade, H.L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, 84, 128-138. <https://doi.org/10.1016/j.tate.2019.05.003>
- Pett, M.A., Lackey, N.R., & Sullivan, J.J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage Publications.
- Price, M., Rust, C., O'Donovan, B., Handley, K., & Bryant, R. (2012). *Assessment literacy: The foundation for improving student learning*. ASKe, Oxford Centre for Staff and Learning Development.
- R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Ramsay, J.O., & Silverman, B.W. (2005). *Functional data analysis* (2nd ed.). Springer.
- Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 2.3.9, <https://CRAN.R-project.org/package=psych>

- Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Rimal, R.N., & Real, K. (2003). Understanding the influence of perceived norms on behaviors. *Communication Theory*, 13(2), 184–203. <https://doi.org/10.1111/j.1468-2885.2003.tb00288.x>
- Sadler, D.R. (2017). Academic achievement standards and quality assurance. *Quality in Higher Education*, 23(2), 81–99. <https://doi.org/10.1080/13538322.2017.1356614>
- Scholl, K., & Olsen, H.M. (2014). Measuring student learning outcomes using the SALG instrument. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, 29(1), 37–50. <https://doi.org/10.1080/1937156X.2014.11949710>
- Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Sage Publications.
- Sijtsma, K., & van der Ark, L.A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, 70(1), 137–158. <https://doi.org/10.1111/bmsp.12078>
- Singh, M., & Ramya, K.R. (2011). Outcome based education. *International Journal of Nursing Education*, 3(2), 87–91.
- Sliter, K.A., & Zickar, M.J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement*, 74(2), 214–226. <https://doi.org/10.1177/0013164413504584>
- Tabachnick, B., & Fidell, L.S. (1996). *Using multivariate statistics*. Harper Collins.
- Van der Ark, L.A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20, 1–19. <https://doi.org/10.18637/jss.v020.i11>
- Van der Ark, L.A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48, 1–27. <https://doi.org/10.18637/jss.v048.i05>
- Velicer, W.F., Eaton, C.A., & Fava, J.L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R.D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 41–71). Kluwer.
- Williams, J. (2016). Quality assurance and quality enhancement: Is there a relationship?. *Quality in Higher Education*, 22(2), 97–102. <https://doi.org/10.1080/13538322.2016.1227207>
- Wolf, R., Zahner, D., & Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: Lessons learned and the road ahead. *Studies in Higher Education*, 40(3), 471–481. <https://doi.org/10.1080/03075079.2015.1004239>
- Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. MESA Press.
- Wright, S.L., & Jenkins-Guarnieri, M.A. (2023). Further validation of the social efficacy and social outcome expectations scale. *Journal of Psychoeducational Assessment*, 42(1), 74–88. <https://doi.org/10.1177/07342829231198277>
- Xu, Y., & Brown, G.T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Zhou, Y. (2019). A mixed methods model of scale development and validation analysis. *Measurement: Interdisciplinary Research and Perspectives*, 17(1), 38–47. <https://doi.org/10.1080/15366367.2018.1479088>
- Zhu, X., & Evans, C. (2022). Enhancing the development and understanding of assessment literacy in higher education. *European Journal of Higher Education*, 1–21. <https://doi.org/10.1080/21568235.2022.2118149>
- Zoom Video Communications, Inc. (2023). *ZOOM cloud meetings* (Version 5.15.5). <https://zoom.com>

An investigation into the effect of different missing data imputation methods on IRT-based differential item functioning

Fatma Ünal^{1*}, Hakan Koğar¹

¹Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye

ARTICLE HISTORY

Received: Jan. 12, 2024

Accepted: Apr. 28, 2024

Keywords:

Differential item functioning,
Missing data,
Item response theory,
Raju's area measurement,
Likelihood ratio.

Abstract: The purpose of this study is to examine the effect of missing data imputation methods, namely regression imputation (RI), multiple imputation (MI) and k-nearest neighbor (kNN) on differential item functioning (DIF). In this regard, the datasets used in the research were created by deleting some of the data via the missing completely at random mechanism from the complete datasets obtained from 600 students in Türkiye, the United Kingdom, the USA, New Zealand and Australia, who answered booklets 14 and 15 from the PISA 2018 science literacy test. Data imputation was applied to the datasets through missing data using RI, MI and kNN methods and DIF analysis was performed on all datasets in terms of language and gender variables via Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method. DIF results from the complete datasets were taken as a reference and they were compared with the results from other datasets. As a result of the research, values close to 10% of accurate imputation were achieved in the RI method depending on language and gender variables. In MI and kNN methods, results closest to the complete datasets were obtained at a rate of 5% depending on the language variable. In the MI method, inaccurate results were obtained in all proportions in terms of the gender variable. For the gender variable, the kNN method gave accurate results at rates of 5% and 10%.

1. INTRODUCTION

Tests developed for the purpose of detecting cognitive or affective characteristics of individuals such as intelligence, achievement, and attitude can be used in many educational studies. According to the scores obtained from the tests used in the field of education, it is possible to examine how much individuals have the characteristics planned to be measured and evaluations can be made based on the results obtained, and important decisions can be raised about individuals (Uyar, 2015; Yılmaz, 2021).

International monitoring studies in education, such as the Program for International Student Assessment (PISA), make it possible for countries to compare their educational status with other countries (MEB, 2019). Thanks to these studies, countries evaluate their education systems and create appropriate policies. PISA is a study conducted in three-year cycles, aimed at evaluating the ability of students aged 15 to reflect the knowledge and skills they have

*CONTACT: Fatma ÜNAL ✉ fatmaunal.452@gmail.com 📧 Akdeniz University, Faculty of Education, Department of Educational Sciences, Antalya, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

acquired in daily life by measuring their science literacy, mathematics literacy and reading skills (MEB, 2019).

Science literacy assesses individuals' ability to engage with science-related topics and scientific phenomena. Individuals who have acquired science literacy should have the ability to explain events in a scientific way, design and evaluate scientific work, stand willing to demonstrate their ability to interpret data and evidence scientifically (OECD, 2019).

There are 3 types of information in science literacy: content, method, and epistemic information. PISA focuses on the capacity of 15-year-old students to reveal these types of information in an appropriate way in personal, local, national, and global situations (OECD, 2019). As a result of the PISA application, the knowledge and skill levels of students in a country can be compared with students in other participating countries. At the same time, standards are established to raise the education levels of countries, and the strengths and weaknesses of education systems can be identified (Taş et al., 2016). Based on this, it can be said that some important inferences can be made about education thanks to studies in education such as PISA. For this reason, to make correct inferences, first, accurate results should be obtained from the studies. Among the reasons for making inaccurate comments and corrections on the research results are the decrease in the validity of the research results and the negative impact on validity. Validity is one of the most important features expected in measurement tools and DIF is one of the factors that cause a decrease in validity (Sırgacı & Çakan, 2020).

In the tests applied in the field of education, individuals at the same ability level are expected to get the same scores from the test items. When individuals in different groups at the same ability level score differently on test items, this indicates that the items are biased towards one group. To determine this bias, differential item functioning analyses are performed on the dataset (Atar et al., 2021, p. 419). DIF analyses assume that the same characteristics of individuals in different subgroups are measured in a test. The goal here is to distinguish between real differences between groups and measurement bias (Kalaycıoğlu & Kelecioğlu, 2011). In order to perform DIF analyses, subgroups are first determined in terms of the variables such as language, gender, and race. The responses to the test items should not differ according to these predetermined subgroups but should differ according to the ability levels of individuals. One of the subgroups is selected as the focus group and the other as the reference group. The responses of the individuals to the test items are compared in the focus group and the reference group. If the probability of answering an item correctly differs from one subgroup to another, it is stated that there is DIF in that item (Dogan et al., 2005). There are some situations that cause DIF in an item. These situations include socio-economic level, comprehensibility of the item, curriculum, poor translation, item writing, the relationship between the content and language of the item and culture, the meaning inferred from the item, and differences in sentence structure (Van de Vijver & Tanzer, 1997). DIF can be analyzed with methods based on item response theory and classical test theory.

Item Response Theory (IRT) consists of a mathematical model indicating the relationship between an individual's observable performance on a test and the latent traits or abilities that are thought to underlie this performance (Hambleton & Swaminathan, 2013, p. 9). With this theory, it is stated that under the assumptions of unidimensionality, local independence, and model-data fit, the estimation of ability parameters can be performed independently of the properties of the items and the estimation of item parameters can be obtained independently of the sample of the study (Gültekin & Demirtaşlı, 2020). In item response theory, the qualifications of the individuals in the study are first determined. Then, scores are estimated for individuals with the relevant qualifications. Thanks to these estimated scores, the test performance of the individual answering the items is determined (Lord & Novick, 2008, p. 359). Item response theory is based on two basic structures:

The latent traits or competencies of individuals can be identified by the performance of respondents on test items.

The relationship between the competencies of the individuals answering the items and their responses to the items can be expressed by a non-linear function called the item characteristic function (Hambleton et al., 1991, p. 110).

The most important difference between item response theory and classical test theory is that in CTT, ability levels are ignored, and a common estimate of measurement precision is used, which is assumed to be equal for all individuals, whereas in IRT, the latent ability value affects measurement precision (Jabrayilov et al., 2016).

To perform DIF analyses based on IRT, unidimensionality and local independence assumptions must be met, and model data fit must be ensured. Unidimensionality is the measurement of a single latent ability of the items included in the test (Hambleton & Swaminathan, 2013, p. 16). Local independence is explained in the form that the item scores of the study group consisting of individuals with the same ability level are independent of each other (Lord & Novick, 2008, p. 361). There are many IRT models available. The widely used unidimensional IRT models are distinguished from each other according to the number of item parameters, and these models are named according to the number of those parameters (Hambleton et al., 1991, p. 12). Logistics models are divided into three: the one-parameter logistic model (1PL), the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). In the one-parameter logistic model (1PL), only the item difficulty parameter is estimated (Hambleton et al., 1991, p. 13). In the two-parameter logistic model (2PL), the item discrimination parameter is estimated in addition to the item difficulty parameter (Hambleton et al., 1991, p. 15). In the three-parameter logistic model (3PL), the chance parameter is estimated in addition to the item discrimination and item difficulty parameters (Hambleton et al., 1991, p. 17).

There are many methods to perform DIF analyses based on IRT. Three of the methods mentioned below were used in this study.

Lord's χ^2 method: In Lord's χ^2 method, the variance and covariance of the focus and reference groups are calculated to detect DIF, and these values are scaled for comparison. These scaled values are calculated using Lord's χ^2 method. At the next stage, the null hypothesis is tested by comparing it with critical values and it is decided whether the difference exists (Cromwell, 2002). According to Lord's χ^2 method, the fact that there is a difference between the focus and reference group item parameters of an item indicates that the item functions in a different way. In other words, for an item to contain DIF, the probabilities of individuals with the same ability level in different groups to respond correctly to the relevant item must differ (Kim et al., 1994).

Raju's area measurement method: In Raju's area measurement method, it is checked whether the area value between the item characteristic curves of two different groups at the same ability level is different from zero, or whether the curves overlap. If the curves overlap, in other words, if the area value measured between the curves is zero, it indicates that the item does not contain DIF (Başusta, 2013). The fact that there is an area between the ICC indicates that the item works differently for the two groups and that the item contains DIF (Raju, 1990).

Item Response Theory Likelihood Ratio method: In this method, the hypothesis of a difference between focus and reference group item parameters is checked. In this respect, restricted and generalized models are created, and the ratios of these models are compared (Atalay et al., 2012). In other words, the significance of the differences in the likelihood ratio values obtained to determine the model-data fit of the restricted and generalized model is tested (Thissen, 2001). The restricted model assumes that the item parameters are the same for the reference and focus groups. In contrast, in the generalized model, it is assumed that the parameters of an item are different for each group while the parameters for the other items are equal. The restricted model is created separately for each item in the study and proportioned to the extended model (Gök et al., 2014).

As with every statistical analysis finding, DIF findings are also affected by the structure and characteristics of the data, such as missing data and outliers.

Missing data is defined as the difference between the data intended to be observed and the observed data (Longford, 2005, p. 13). There are many reasons for the occurrence of missing data. For example, missing data may exist due to some individuals in the sample not answering the questions unconsciously, participants preferring not to answer the questions, participants leaving the study before it is completed, problems arising during data collection or problems arising from the data collection tool, and due to errors made during data entry (Osborne, 2013, pp. 106-108).

Missing data can cause some problems: It can create bias in the estimations in statistical analyses, reduce the power of the analysis and lead to higher standard error values due to lack of information. Furthermore, frequently used statistical methods cannot be applied to datasets with missing data leading to improper use of assessable resources (Peng et al., 2006).

In order to make accurate predictions in research, a solution to the problem of missing data should be found before proceeding with the analysis. In this direction, researches may consider solutions such as including new values in the data, not including cells with missing data in the dataset, making predictions about missing data and imputing approximate values instead of missing data (Çüm & Gelbal, 2015).

To impute values to missing data, it is necessary to choose the appropriate imputation method. For this purpose, firstly, the structure of the missing data is examined, and the appropriate imputation method is selected. Missing data can occur in three different mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR):

MCAR is defined as missing data that is not affected by the variable in which the missing data is located and is not caused by other variables such as language or gender (Çüm et al., 2018). For example, if the missing values in a dataset consisting of students' answers to exam questions do not differ for students with high or low scores, or if any other variable did not have an effect on the missing values, it can be said that the missing data are distributed completely at random (MCAR).

MAR means that the missing data for a variable are not caused by that variable but by the effect of one or more other variables in the model (Enders, 2010, p. 6). For example, the fact that the missing data in the variable consisting of students' answers to the exam questions do not differ according to the high or low scores obtained, but the effect of one or more other variables on the losses shows that the missing data are randomly distributed.

MNAR is defined as the probability that having missing data in a variable is related to the values of the relevant variable even after controlling other variables. In other words, the probability of missing data affects the variable with missing values (Enders, 2010, p. 8). For example, the fact that the missing values obtained from the variable including students' answers to the exam questions differ for individuals with high or low scores but are not affected by other variables shows that missing data are distributed not at random (MNAR).

There are methods suitable for missing data mechanisms. Among these methods, the ones used in the research are explained below:

Regression Imputation (RI) Method: In the regression imputation method, a regression equation is first established that predicts the missing data from the complete data. Then, estimated values are created, and these values are substituted for the missing data to obtain a complete dataset (Enders, 2010, p. 44). Regression imputation provides unbiased parameter estimates in MCAR and MAR missing data mechanisms (Baraldi & Enders, 2010). This method has some negative features: Since the missing data are estimated based on the complete data, results close to the other data will be obtained. Therefore, results similar to the real data will not be obtained. And

the variance will decrease because the data obtained by regression imputation make predictions close to the average. When the independent variables are not good, this method will reach the same results as the mean imputation method because it will not be able to predict the missing data accurately. Finally, this method cannot be used when the value obtained with the regression imputation method is not within the data value range (Tabachnick & Fidell, 2013, p. 68).

Multiple imputation (MI) Method: In this method, the missing data imputation process takes place in three steps. In the first step, m ($m > 1$) complete datasets are created. In the second step, m different datasets are analysed with standard methods. Finally, the results of the analyses are combined to form a single dataset (Schafer & Graham, 2002). In this method, missing data imputation is iterated at least 2 times and there is no limit to the number of iterations. A large number of imputations with the MI method reduces the standard error (Schafer & Olsen, 1998). This method makes accurate inferences even in MAR and MNAR mechanisms (Van Buuren, 2018, p. 48).

K-Nearest Neighbor Method (kNN): In this method, data imputation is performed by distance-based classification (Cihan, 2018). The kNN method imputes missing data in four stages. In the first stage, the distances between the target data and other data are calculated. In the second stage, these distances are ranked, and in the third stage, the k smallest values between the ranked distances are taken. In the last stage, the target data is imputed to the most repeated class among the k values (Altay, 2016). The characteristics of all groups should be identified in advance. The effectiveness of the k -nearest neighbor method is affected by some conditions. The number of neighbors, threshold value, similarity measurement and sufficient number of normal actions in the learning set can be given as examples (Çalışkan & Soğukpınar, 2008).

Like many statistical methods, DIF analyses are also affected by the existence of missing data since they are developed for complete datasets. Therefore, if there is missing data in the dataset, the missing data problem should be solved with appropriate methods and the dataset should be made complete before proceeding to DIF analysis. A review of the literature reveals that there are few studies on the effect of missing data imputation methods on DIF. In the studies examining the effect of missing data imputation methods on DIF, it has been found out that DIF methods based on CTT are generally used or DIF methods based on CTT are compared with DIF methods based on IRT (Dinçsoy, 2022; Emenogu et al., 2010; Garrett, 2009; Robitzsch & Rupp, 2009; Selvi & Alici, 2018; Tamcı, 2018). The fact that DIF methods based on IRT are not generally used in the studies revealed that conducting a study on DIF methods based on IRT would contribute to the literature. At the same time, because of this study, it is aimed to enable the selection of appropriate imputation methods for future IRT-based DIF analyses. Based on this objective, this study examines the effects of regression imputation (RI), multiple imputation (MI), k -nearest neighbor (kNN) methods on DIF detection through Lord's χ^2 , Raju's area measurement, item response theory likelihood ratio methods.

2. METHOD

2.1. Research Model

This research aims to examine the effect of regression imputation (RI), multiple imputation (MI), and k -nearest neighbor (kNN) methods on DDIF to deal with missing data in a dataset containing missing values at different rates considering the variables of language and gender using Lord's χ^2 , Raju's area measurement and item response theory likelihood ratio methods. For this reason, a descriptive survey model was used in the study. The descriptive survey model examines existing phenomena in terms of conditions, practices, beliefs, processes, relationships, or trends (Salaria, 2012).

2.2. Study Group

International studies such as the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Cooperation and Development (OECD) allow

comparing the performance of students in different countries (Taş et al., 2016). To carry out DIF analyses in terms of the language variable in the study, PISA 2018 data from different countries were used in this study. Accordingly, the study group of the research consists of students who answered the PISA 2018 science literacy test. 600.000 students participated in PISA 2018, representing approximately 32 million students in the 15-age group from 79 participating countries and economies (OECD, 2019). For the study, Türkiye and the countries that use English as their mother tongue, one of the languages in which the PISA 2018 tests were developed, were selected to conduct DIF analyses for the language variable. In the selection of the countries whose mother tongue is English, attention was paid to pick the ones with the closest science averages to each other. For this reason, the United Kingdom, the United States, New Zealand, and Australia were included in this study. In PISA 2018, 6.890 students from Türkiye, 13.818 students from the United Kingdom, 4.838 students from the United States, 6.173 students from New Zealand, and 14.273 students from Australia participated (OECD, 2019).

Sample sizes with equal focal and reference groups reduce the error rates of DIF detection methods based on IRT (Sünbül & Sünbül, 2016). Thus, as many native English speakers as the number of native Turkish speakers were included in the analysis through simple random sampling. There were 530 students from Türkiye who participated in PISA 2018 answering booklets 14 and 15. Since 300 students out of 530 students had the complete responds, those 300 students from Türkiye were included in the study. Accordingly, out of 1.756 students from native English-speaking countries who answered the items in booklets 14 and 15 and had complete respondst, 300 of them were chosen for the analysis by simple random sampling method. A total of 600 students from Türkiye, the United Kingdom, the United States, New Zealand, and Australia were included in the analysis. When the literature was examined, it was seen that the sample size of the focus and reference groups should be larger than 200 in the analyses related to DIF because it is important in obtaining accurate results (Jodoin & Gierl, 2001; Rogers & Swaminathan, 1993). Based on this, it can be stated that accurate DIF results can be obtained from the analyses when the sample is examined. [Table 1](#) shows the distribution of individuals in the study group by country and gender with science test means of countries.

Table 1. *Distribution of individuals in the study group by country and gender and science means of countries.*

Country	Female	Male	Total	Mean Science Literacy
Australia	41	70	121	503
United Kingdom	37	53	90	505
New Zealand	18	23	41	508
Türkiye	147	153	300	468
USA	22	26	48	502
Total	283	317	600	

When [Table 1](#) is examined, there are 121 students (41 female and 70 male) from Australia in the dataset. There are 90 students (37 female and 53 male) from the United Kingdom; 41 students (18 female and 23 male) from New Zealand; 300 students (147 female and 153 male) from Türkiye, and 48 students (22 female and 26 male) from the USA. The dataset of the study consisted of 600 students comprising 283 females and 317 males. When the mean science literacy scores of the countries are analysed, Australia has a mean score of 503, the United Kingdom 505, New Zealand 508, Türkiye 468, and the USA 502.

2.3. Data Collection Tools

This study was conducted on booklets numbered 14 and 15, which have the highest number of common items among the booklets used for Turkish and English languages in the PISA 2018 application and which provide content validity. The booklets included in the study had a total of 20 common items, 5 open-ended and 15 multiple-choice items. Correct answers were coded as “1” and incorrect answers were coded as “0”. In the items with partially correct answers, incorrect answers were coded as “1”, partially correct answers as “11” and “12”, and correct answers as “21”. The answers with the codes “5, 6, 7, 8, 9, 96, 97, 98, 99” were included in the analysis with the missing data code as in the PISA 2018 codebook. In the item numbered DS657Q04C with partially correct answers, answers coded “21” were coded as “1”; answers coded “1”, “11” and “12” were coded as “0” and included in the analysis. The data for PISA 2018 were published on the OECD website in 2019 (<https://www.oecd.org/pisa/data/>). Table 2 provides information about the items included in the study.

Table 2. Science literacy items used in the analysis.

Item	Unit	Scientific competencies	Content
CS466Q01S	Forest Fires	Evaluate and design scientific enquiry	Physical
CS466Q07S	Forest Fires	Evaluate and design scientific enquiry	Physical
CS256Q01S	Spoons	Explain phenomena scientifically	Physical
DS326Q01C	Milk	Interpret data and evidence scientifically	Living
DS326Q02C	Milk	Interpret data and evidence scientifically	Living
CS326Q03S	Milk	Interpret data and evidence scientifically	Living
CS326Q04S	Milk	Interpret data and evidence scientifically	Physical
CS602Q01S	Urban Heat Island Effect	Interpret data and evidence scientifically	Earth and Space
CS602Q02S	Urban Heat Island Effect	Explain phenomena scientifically	Earth and Space
DS602Q03C	Urban Heat Island Effect	Explain phenomena scientifically	Physical
CS602Q04S	Urban Heat Island Effect	Interpret data and evidence scientifically	Living
CS603Q03S	Elephants and Acacia Trees	Explain phenomena scientifically	Living
DS603Q02C	Elephants and Acacia Trees	Evaluate and design scientific enquiry	Living
CS603Q03S	Elephants and Acacia Trees	Explain phenomena scientifically	Living
CS603Q03S	Elephants and Acacia Trees	Explain phenomena scientifically	Living
CS603Q05S	Elephants and Acacia Trees	Evaluate and design scientific enquiry	Living
CS657Q01S	Invasive Species	Explain phenomena scientifically	Living
CS657Q02S	Invasive Species	Explain phenomena scientifically	Living
CS657Q03S	Invasive Species	Interpret data and evidence scientifically	Living
DS657Q04C	Invasive Species	Explain phenomena scientifically	Living

When Table 2 is examined, it can be observed that the PISA 2018 science literacy test items included in the study are found in the units of forest fires, spoons, milk, urban heat island effect, elephants and acacia trees, and invasive species. The items measure the skills of evaluating and designing scientific research, explaining phenomena scientifically, and interpreting data and evidence scientifically. Physical, living, Earth and Space titles constitute the content areas of the items.

2.4. Data Analysis

In the study, outliers and descriptive statistics were checked first via the IBM SPSS 26.0 program. Then, confirmatory factor analysis was conducted with the “lavaan” package of the R Studio program to test the unidimensionality and local independence assumptions regarding IRT (Rossee et al., 2017). R Studio program “ltm” package was used to examine model-data fit (Rizopoulos & Rizopoulos, 2018). The population heterogeneity of the dataset was examined with the “Equaltest MI” package of the R Studio program (Jiang et al., 2022). After reviewing the suitability of the dataset for analysis, four datasets with 5%, 10%, 20% and 30% of missing data suitable for the MCAR mechanism were created from the complete dataset with the R Studio program “MissMethods” package (Josse et al., 2022) and the missing data mechanisms of the datasets were checked with the IBM SPSS 26.0 program. In the following stage, the missing data were imputed via the RI and MI methods using the IBM SPSS 26.0 program and the kNN method using the R Studio program “VIM” (Templ et al., 2016) package. With the MI method, imputations were made by selecting 5 as the imputation number and 5 different datasets belonging to each missing data rate were obtained. For each dataset with missing data rates in the study, the average of the DIF analyses of the 5 imputations made with the MI method were combined in a common DIF result. DIF analyses were performed with Lord’s χ^2 , Raju’s area measurement and item response theory likelihood ratio methods using the R Studio program “difR” (Magis et al., 2015) package in terms of gender and language variables for the datasets completed by imputations via RI, MI, kNN methods. The values obtained from the complete datasets were taken as a reference and compared with the results obtained from the datasets in which missing data were imputed.

2.4.1. Outliers

Outliers are explained as data with values outside the usual values or extreme values (Çokluk et al., 2021, p. 2). Outliers can occur in two ways: univariate and multivariate. Univariate outliers can be detected by statistical methods such as converting all raw scores in the distribution into standard Z scores. For a subject to be an outlier, the Z value must be less than -3 and greater than +3 (Çokluk et al., 2021, p.14). To detect the univariate outliers in the dataset, the Z values were examined. As a result of the analysis conducted to detect the Z value, it was found that there are no univariate outliers in the dataset since a Z value less than -3 and greater than +3 was not detected. To determine the multivariate outliers, Mahalanobis Distance, which measures a single data distance from the center or sample mean in the space of the independent variable, is used. A Mahalanobis Distance value of $p < 0.001$ indicates that multivariate outliers are present in the dataset (Çokluk et al., 2021, p.15). When the Mahalanobis Distance was examined for the dataset, the data with a value less than 0.001 could not be determined and it was seen that the multivariate outliers were not present in the dataset.

2.4.2. Descriptive test statistics

Some statistical options such as kurtosis and skewness coefficients can be used to assess the normality of the dataset. Skewness and kurtosis coefficients between +1 and -1 indicate that the group does not deviate excessively from the normal distribution (Çokluk et al., 2021, p. 16).

In this study, internal consistency was tested by examining the Kuder Richardson-20 (KR-20) coefficient. A KR-20 reliability coefficient of 0.70 and above indicates that the internal consistency value is at an acceptable level (De Vellis, 2003, p. 95).

In this section, the normality of the data was examined. [Table 3](#) presents the findings related to the normality and reliability tests.

Table 3. Test statistics, normality tests and reliability coefficients related to sub-problems.

Statistics	Gender		Language	
	Female	Male	Turkish	English
Number of Students	283	317	300	300
Mean	11.6	11.82	11.02	12
Median	12	12	11	13
Mode	9	15	11	15
Standard Deviation	3.96	4.11	3.97	3.99
Range	18	20	18	20
Skewness	-0.31	-0.36	-0.08	-0.63
Kurtosis	-0.60	-0.55	-0.63	-0.19
KR-20	0.77	0.79	0.76	0.79

When examining [Table 3](#), it is evident that the measures of central tendency are relatively close to each other. Skewness and kurtosis coefficients are in the range of +1 and -1. This indicates that the distribution is close to normal (Çokluk et al., 2021, p. 16). The KR-20 reliability coefficients of 0.70 and above in all groups indicate that the reliability principle of the groups is met.

2.4.3. Confirmatory factor analysis

In this study, confirmatory factor analysis was performed on the complete dataset with the R Studio program “lavaan” package to examine whether the data has met the unidimensionality assumption (Rosseel et al., 2017).

Table 4. Confirmatory factor analysis model data fit values.

Indices	Value
$SB\chi^2$	222.31
Degrees of freedom	167
RMSEA	0.02
SRMR	0.03
TLI	0.94
CFI	0.95

As a result of confirmatory factor analysis, $SB\chi^2$, degrees of freedom, RMSEA, SRMR, TLI, and the CFI values were obtained and the unidimensionality assumption was checked based on these values. The Tucker and Lewis index (TLI) value above 0.97 indicates perfect fit, above 0.95 indicates very good fit, and above 0.85 indicates acceptable fit. The standardized root mean square of residuals (SRMR) values close to 0 are considered excellent and values less than 0.05 are considered good. The root mean square error of approximation (RMSEA) value is considered good when it is 0.05 and less, acceptable between 0.05 and 0.08, and poor when it is 0.10 and above. The comparative fit index (CFI) shows an acceptable fit between 0.95 and 0.97 (Erdoğan, 2019). Based on this information, when [Table 4](#) created as a result of confirmatory factor analysis is examined, it is determined that all values provide model-data fit. This shows that the unidimensionality assumption is met.

Local independence is an assumption related to the unidimensionality assumption. If the unidimensionality assumption is met in a test, the items in the test also meet the local independence assumption (Hambleton & Swaminathan, 2013, p. 23). Accordingly, it can be stated that the items in the study meet the local independence assumption.

2.4.4. Population heterogeneity

In this study, to determine the suitability of the dataset for the analysis, the population heterogeneity of the dataset was checked in terms of language and gender variables using the "Equaltest MI" package of the R Studio program (Jiang et al., 2022). To determine population heterogeneity, Model 5 and Model 6 were compared for equality in latent means. S-B $\chi^2(df)$, χ^2/df , $\Delta\chi^2(\Delta df)$, RMSEA, Δ RMSEA goodness-of-fit indices of the models were taken into account during the comparison. A value range of $0 \leq \chi^2/df \leq 2$ indicates a good fit and a value range of $2 \leq \chi^2/df \leq 3$ indicates an acceptable fit. While a value range of $0 \leq \text{RMSEA} \leq 0.05$ indicates a good fit, and a value range of $0.05 \leq \text{RMSEA} \leq 0.08$ indicates an acceptable fit (Schermelleh-Engel et al., 2003). In this study, the change between Model 5 and Model 6 was evaluated by considering $\Delta\text{CFI} \leq 0.01$ and $\Delta\text{RMSEA} \leq 0.01$ (Taşkıran & Şenel, 2022).

Table 5. Population heterogeneity fit indices of the dataset by language and gender variables.

	Model	S-B $\chi^2(df)$	χ^2/df	$\Delta\chi^2(\Delta df)$	CFI	ΔCFI	RMSEA	ΔRMSEA
Language	Model 5	618.41 (388)	1.75		0.77		0.05	
	Model 6	702.62 (391)	1.80	21.21(3)	0.75	0.01	0.05	0.00
Gender	Model 5	455.78 (388)	1.17		0.95		0.02	
	Model 6	458.91 (391)	1.17	3.12(3)	0.95	0.00	0.02	0.00

$p < 0.05$, Model 5 = Equality of variance, Model 6 = Equality of Latent Means

When the χ^2/df indexes are examined in terms of the language variable in Table 5, the fact that Model 5 has a value of 1.75 and Model 6 has a value of 1.80 indicates that both models show a good fit. The ΔRMSEA value of 0 indicates that this fit index is at an acceptable level. Based on this, it can be said that there is a good fit between the models. When the ΔCFI fit index is examined, the fact that this value is 0.01 indicates that the fit index is at an acceptable level proving that there is a good fit between the models.

Considering the χ^2/df index in terms of the gender variable, Model 5 and Model 6 have a value of 1.17 indicating a good fit. ΔRMSEA value of 0 indicates that the fit index is at an acceptable level and there is a good fit between the models. A ΔCFI value of 0 indicates that the fit index is at an acceptable level and there is a good fit between the models. According to the results of the population heterogeneity analysis, it was determined that there was no difference between the latent means for both variables.

2.4.5. Model-data fit

In this study, model-data fit was examined through the "lrm" package in the R Studio program (Rizo-Poulos & Rizopoulos, 2018). For this reason, the likelihood ratio test (logLik), Akaike information criterion (AIC) and Bayesian information criterion (BIC) values were compared, and the p -value and degrees of freedom obtained as a result of ANOVA were analyzed. Table 6 shows the results of the model-data fit analysis.

Table 6. Model data fit comparison.

Model	logLik	AIC	BIC	Number of Parameters	degrees of freedom	p
Rasch-1PL	-6681.39	13404.78	13497.11	14		
2PL	-6632.76	13345.53	13521.40	14	19	0
3PL	-6620.33	13360.67	13624.48	16	20	0.20

When Table 6 is examined, the fact that the p -value of the 3PL model is not significant ($p > 0.05$) indicates that the model is not suitable for analysis. The fact that the loglik and AIC values of the 2PL model are smaller than the loglik and AIC values of the 1PL model indicates that the 2PL model is suitable for the study. Although the fact that the BIC value of the 1PL model is

less than the 2PL model does not support this situation, the fact that the variance analysis result of the 2PL model is significant shows that 2PL model fits better than other models and as a result, the 2PL model is the appropriate model for the analysis. In the study, after factor analysis, population heterogeneity and model-data fits were examined, four datasets with 5%, 10%, 20% and 30% of missing data suitable for the MCAR mechanism were created from the complete dataset and the missing data mechanisms of the datasets were checked. In the next stage, the datasets were completed by imputing missing data using RI, MI and kNN methods. DIF analyses were performed on the newly obtained datasets with gender and language variables using Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method. As a result of the analyses, items with a p -value below 0.05 and DIF finding in two of the three DIF methods were accepted to contain DIF. Accordingly, DIF analyses were performed on the complete dataset and datasets with missing data imputation. The values obtained from the complete datasets were compared with the results obtained by data imputation.

3. FINDINGS

In this section, the results of the DIF analyses are presented. In the analyses, Lord's χ^2 method, Raju's area measurement method, and item response theory likelihood ratio method are applied for gender and language variables. The analyses were carried out on the complete dataset and the one with missing data. The missing dataset was completed by imputing 5%, 10%, 20%, and 30% via RI, MI, and kNN methods. In [Table 7](#), the DIF results obtained by Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method from the complete dataset and the datasets completed by imputing 5%, 10%, 20% and 30% in terms of the language variable and [Table 8](#) in terms of the gender variable are compared. If at least two of the three DIF detection methods used in the study showed DIF, the related item was considered to contain DIF. In [Table 7](#) and [Table 8](#), in the complete dataset and in the datasets completed with RI, MI, and kNN methods at the rates of 5%, 10%, 20%, and 30%, "DIF" was written in front of the items that showed DIF in at least two DIF detection methods and it was stated that the relevant item contained DIF.

In [Table 7](#), the items in the complete dataset and the datasets completed with RI, MI and kNN methods at 5%, 10%, 20% and 30% of rates were identified as DIF items in terms of the language variable using Lord's χ^2 method, Raju's area measurement method, and item response theory likelihood ratio method. If DIF was identified in at least two methods among the items in the datasets, it was accepted that the item contained DIF. Accordingly, DIF was detected in 6 items (CS256Q01S, CS326Q04S, CS602Q01S, CS603Q01S, DS603Q02C, CS603Q03S) out of 20 items included in the analysis in the complete dataset.

DIF was detected in 6 items (CS256Q01S, CS326Q04S, CS602Q01S, CS602Q02S, CS603Q03S, CS603Q04S) in the dataset that was imputed at 5% with the RI method. There was a 67% agreement between the complete dataset and the dataset completed by 5% with the RI method regarding items containing DIF.

In the dataset completed 10% by the RI method, DIF was detected in 5 items (CS256Q01S, CS326Q04S, CS602Q01S, S603Q02C, CS603Q03S). Based on this, 83% agreement was found between the items with DIF in the complete dataset and those with DIF in the dataset completed 10% with the RI method.

In the dataset with 20% missing data imputation by the RI method, DIF was found in 3 items (DS603Q02C, CS603Q03S, CS603Q04S). Between the complete dataset and the dataset completed by the RI method at the rate of 20%, the rate of the same items containing DIF was determined as 33%.

In the dataset completed 30% with the RI method, DIF was found in 2 items (CS603Q03S, CS603Q05S). The probability of the same items containing DIF was found to be 17% in the dataset in which 30% of the data were imputed by the RI method.

DIF was detected in 6 items (CS326Q04S, CS602Q01S, CS603Q01S, DS603Q02C, CS603Q03S, CS603Q04S) in the dataset completed 5% with the MI method. It was observed that 83% of the items with DIF in the complete dataset also contained DIF in the one completed 5% with the MI method.

Table 7. Findings of item response theory-based differential item functioning (Lord’s χ^2 , Raju’s area measurement, item response theory likelihood ratio) analysis of complete dataset and datasets with different ratios of missing data and completed with different imputation methods (regression imputation, multiple imputation and k-nearest neighbor method) in terms of the language variable.

Item	DIF Status												
	complete dataset	RI Method				MI Method				kNN Method			
		5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
CS466Q01S													
CS466Q07S													
CS256Q01S	DIF	DIF	DIF							DIF			
DS326Q01C								DIF					
DS326Q02C									DIF				
CS326Q03S									DIF			DIF	
CS326Q04S	DIF	DIF	DIF			DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF
CS602Q01S	DIF	DIF	DIF			DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF
CS602Q02S		DIF											
DS602Q03C													
CS602Q04S													
CS603Q01S	DIF					DIF							
DS603Q02C	DIF		DIF	DIF		DIF		DIF		DIF		DIF	
CS603Q03S	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF
CS603Q04S		DIF		DIF		DIF						DIF	DIF
CS603Q05S					DIF								
CS657Q01S												DIF	
CS657Q02S													DIF
CS657Q03S													
DS657Q04C									DIF				

DIF was detected in 3 items (CS326Q04S, CS602Q01S, CS603Q03S) in the dataset that was made complete by imputing 10% of data with the MI method. 50% of the items with DIF in the complete dataset also showed DIF in the dataset with 10% of data imputation by the MI method.

In the dataset, where 20% of the missing data was imputed with the MI method, DIF was observed in 5 items (DS326Q01C, CS326Q04S, CS602Q01S, DS603Q02C, CS603Q03S). 67% of the items detected DIF in the complete dataset contain DIF in the dataset with 20% of data imputation by the MI method.

In the dataset completed 30% with the MI method, DIF was detected in 6 items, including items numbered DS326Q02C, CS326Q03S, CS326Q04S, CS602Q01S, CS603Q03S, DS657Q04C. 50% of the items containing DIF in the complete dataset also contain DIF in the one completed 30% with the MI method.

It was observed that 5 items (CS256Q01S, CS326Q04S, CS602Q01S, DS603Q02C, CS603Q03S) contained DIF in the dataset with 5% of imputation by the kNN method. It was found that the items containing DIF in the dataset completed by the kNN method at the rate of 5% were the same items as 83% of the items detected DIF in the complete dataset.

In the dataset completed 10% applying the kNN method, DIF was detected in 3 items: CS326Q04S, CS602Q01S, and CS603Q03S. 50% of the items with DIF in the full data set also showed DIF in the data set where 10% were assigned by the MP method.

In the dataset with 20% missing data imputation by the kNN method, DIF was detected in 7 items (C6S326Q03S, CS326Q04S, CS602Q01S, DS603Q02C, CS603Q03S, CS603Q04S, CS657Q01S). The ratio of the number of common items between the items containing DIF in the complete dataset and the items containing DIF in the dataset in which 20% of the data was imputed with the kNN method is 67%.

5 items (CS326Q04S, CS602Q01S, CS603Q03S, CS603Q04S, CS657Q02S) contain DIF in the dataset completed by imputing 30% with the kNN method. 50% of the items with DIF in the complete dataset are compatible with the dataset made 30% complete by the kNN method.

Table 8. Findings of item response theory-based differential item functioning (Lord’s χ^2 , Raju’s area measurement, item response theory likelihood ratio) analysis of complete dataset and datasets with different ratios of missing data and completed with different imputation methods (regression imputation, multiple imputation and k-nearest neighbor method) in terms of the gender variable.

Item	DIF Status												
	complete dataset	RI Method				MI Method				KNN Method			
		5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%
CS466Q01S													
CS466Q07S													
CS256Q01S													
DS326Q01C													
DS326Q02C				DIF								DIF	
CS326Q03S												DIF	
CS326Q04S													
CS602Q01S													
CS602Q02S													
DS602Q03C													
CS602Q04S													
CS603Q01S	DIF		DIF							DIF	DIF		
DS603Q02C													
CS603Q03S													
CS603Q04S													
CS603Q05S													
CS657Q01S													
CS657Q02S													
CS657Q03S													

In Table 8, items showing DIF in terms of the gender variable were identified through Lord’s χ^2 method, Raju’s area measurement method, and item response theory likelihood ratio method from the items in the complete dataset and the datasets completed with RI, MI and kNN methods at 5%, 10%, 20% and 30% of rates. If DIF was detected in at least two methods, it was accepted that the item contained DIF. Based on this, DIF was found in the item DS603Q01S included in the analysis of the complete dataset.

DIF could not be determined in any item in the datasets completed by imputing 5% and 30% of missing data using the RI method. This shows that the DIF inclusion rate of the same items is 0% between the datasets with 5% and 30% of data imputation using the RI method and the complete dataset.

The detection of DIF in the item CS603Q01S in the dataset completed at the rate of 10% by RI shows that the same item contains DIF both in the complete dataset and the dataset imputed 10% by the RI method. DIF inclusion rate of the same items is 100% between the complete dataset and the one with %10 data imputation using the RI method.

In the dataset, completed by imputing the missing data by the RI method at the rate of 20%, DIF was found in item DS326Q02C. This shows that the DIF inclusion rate of the same items is 0% between the dataset with 20% data imputation using the RI method and the complete dataset.

DIF could not be determined in any item completed 5%, 10%, 20% and 30% by the MI method. The fact that there are no items containing DIF in the datasets completed 5%, 10%, 20% and 30% with the MI method indicates that the DIF rate of the complete dataset and these datasets is 0% for the same items.

In the datasets, completed by imputing 5% and 10% by the kNN method, it was found that the item CS603Q01S contained DIF. DIF inclusion rate of the same items is 100% between the datasets with 5% and 10% data imputation using the kNN method and the complete dataset.

DIF was detected in 2 items (DS326Q02C, C6S326Q03S) in the dataset completed by the kNN method at the rate of 20%. The DIF rate of the same items is 0% between the complete dataset and the dataset with %20 data imputation using the kNN method.

It was determined that DIF was not observed in any item in the dataset in which 30% of the missing data were imputed by the kNN method and that different DIF findings were obtained with the complete dataset. There was a 0% agreement between the complete dataset and the dataset completed 30% by the kNN method.

4. DISCUSSION and CONCLUSION

In this study, we examined how DIF results obtained with Lord's χ^2 , Raju's area measurement and item response theory likelihood ratio methods change according to the missing data rate using the language and gender variables in the datasets completed by imputing 5%, 10%, 20% and 30% of data using RI, MI and kNN methods. In this regard, the study was conducted on PISA 2018 science literacy test items.

As a result of the analyses, it can be stated that the use of different languages by the individuals responding to the relevant items increases the probability of the items containing DIF because DIF was observed in 6 out of 20 items in the complete dataset regarding the language variable. Observing DIF in 1 out of 20 items in terms of the gender variable in the complete dataset, it can be said that the gender of individuals affects the probability of DIF. With the RI method, the closest result to the complete dataset using the language variable was obtained at a rate of 10%. While better results were obtained at 5% compared to 20% and 30%, the worst result was obtained at 30%. By the gender variable in the completed datasets with the RI method, accurate results were obtained at 10%, while inaccurate results were obtained at 5%, 20% and 30%. In the MI method, the closest result to the complete dataset was obtained at 5% in terms of the language variable while more accurate predictions were made at 20% compared to 10% and 30%. Tamcı (2019) suggested that the MI method should be used when the missing data rate is high. Dıngsoy (2022) found that the MI method was successful in detecting DIF at 10% and 20% of missing data. In the MI method, inaccurate results were obtained at 5%, 10%, 20% and 30% with the gender variable. With the kNN method, values close to the complete dataset were obtained at 5% by the language variable while the most accurate results were obtained after 5% at 20%. DIF was poorly predicted at 10% and 30% rates compared to other rates. The kNN

method obtained accurate results at 5% and 10% of missing data rates regarding the gender variable, but inaccurate results were obtained at 20% and 30% of rates. Based on the results of the study, it can be said that the RI method can be used to make imputations at a 10% missing data rate in future studies analyzing DIF based on IRT by the variables of language and gender. It can be suggested that the RI method should not be used at 5%, 20% and 30% of missing data rates. In terms of the language variable, it can be recommended that MI and kNN methods can be used at a rate of 5% in DIF analysis based on IRT, but these methods should not be used at 10%, 20% and 30% of missing data rates. Since inaccurate results will be obtained with the MI method at 5%, 10%, 20% and 30% by the gender variable, it may be recommended to prefer different missing data imputation methods. It can be suggested that the kNN method can be used in the dataset with 5% and 10% of missing data for the gender variable, but this method should not be preferred at 20% and 30% rates. Since the sample size was kept constant in this study, missing data imputation methods with different sample sizes can be examined in future studies. In this study, missing data with the MCAR mechanism were used. In future studies, DIF analyses can be performed with missing data with MAR and MNAR mechanisms.

There are some limitations in this study. It is limited to the use of regression imputation, multiple imputation and k-nearest neighbor imputation methods, and IRT-based Lord's χ^2 method, Raju's area measurement method and item response theory likelihood ratio method for DIF identification. Therefore, in future studies, different DIF detection methods based on IRT or CTT, different missing data imputation methods, and the effect of those imputation methods on DIF can be examined.

Acknowledgments

This study is produced from the first author's master's dissertation.

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

Fatma Ünal: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Hakan Koğar:** Methodology, Supervision, and Validation.

Orcid

Fatma Ünal  <https://orcid.org/0000-0001-6306-4210>

Hakan Koğar  <https://orcid.org/0000-0001-5749-9824>

REFERENCES

- Altay, O. (2016). *Genetik ve genetik olmayan faktörlere bağlı olarak Türk hastalarda varfarin dozajını tahmin eden bir uzman sistem geliştirilmesi [Improvement of an expert system that predict warfarin dosage in Turkish patients depending on genetic and non-genetic factors]* [Master's dissertation, Fırat University]. Higher Education Institution National Thesis Center. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=W663t01X1WehurHffLL0Q&no=Urx32Vn-YC2f6ufE0L3ZTW>
- Atalay, K., Gök, B., Kelecioğlu, H., & Arsan, N. (2012). Değişen madde fonksiyonunun belirlenmesinde kullanılan farklı yöntemlerin karşılaştırılması: Bir simülasyon çalışması [Comparing different differential item functioning Methods: A simulation study]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education)*, 43, 270-281. <https://dergipark.org.tr/pub/hunefd/issue/7795/102030>

- Atar, B., Atalay Kabasakal, K., Ünsal Özberk, E.B., Özberk, E.H., & Kıbrıslıoğlu Uysal, N., (2021). *R ile veri analizi ve psikometri uygulamaları [Data analysis and psychometric applications with R]* (3th ed.). Pegem Akademi.
- Baraldi, A.N., & Enders, C.K. (2010). An introduction to modern missing data analyses. *Journal of school psychology, 48*(1), 5-37. <https://doi.org/10.1016/j.jsp.2009.10.001>
- Başusta, N.B. (2013). *PISA 2006 fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi [Examination of item bias and language perspective of PISA 2006 science and culture achievement test]* [Doctoral dissertation, Hacettepe University]. Hacettepe University Open Archive. <https://www.openaccess.hacettepe.edu.tr/xmlui/bitstream/handle/11655/1766/42cc60c5-40f1-4b78-8c75-cc6d7932416e.pdf?sequence=1&isAllowed=y>
- Bortolotti, S.L.V., Tezza, R., de Andrade, D.F., Bornia, A.C., & de Sousa Júnior, A.F. (2013). Relevance and advantages of using the item response theory. *Quality & Quantity, 47*, 2341-2360.
- Cihan, P. (2018). *Veri madenciliği yöntemleriyle hayvan hastalıklarında teşhis, prognoz ve risk faktörlerinin belirlenmesi [Determination of diagnosis, prognosis and risk factors in animal diseases using by diseases using by data mining methods]* [Doctoral dissertation, Yıldız Technical University]. Yıldız Technical University Open Archive. <http://dspace.yildiz.edu.tr/xmlui/bitstream/handle/1/13155/7932.pdf?sequence=1&isAllowed=y>
- Cromwell, S. (2002). A primer on ways to explore item bias. <https://eric.ed.gov/?id=ED463307>
- Çalışkan, S.K., & Soğukpınar, İ. (2008). Kxknn: K-means ve k en yakın komşu yöntemleri ile ağlarda nüfuz tespiti [Kxknn: Penetration detection in networks with k-means and k nearest neighbor methods]. *EMO Yayınları, 120-24*. https://www.emo.org.tr/ekler/8c1874c96244659_ek.pdf
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2021). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]* (6th ed.). Pegem Akademi.
- Çüm, S., & Gelbal, S. (2015). Kayıp veriler yerine yaklaşık değer atamada kullanılan farklı yöntemlerin model veri uyumu üzerindeki etkisi [The effects of different methods used for value imputation instead of missing values on model data fit statistics]. *Mehmet Akif Ersoy University Journal of Education Faculty, 1*(35), 87-111. <https://dergipark.org.tr/tr/pub/maeuefd/issue/19408/206357>
- Çüm, S., Demir, E.K., Gelbal, S., & Kışla, T. (2018). Kayıp veriler yerine yaklaşık değer atamak için kullanılan gelişmiş yöntemlerin farklı koşullar altında karşılaştırılması [A comparison of advanced methods used for missing data imputation under different conditions]. *Mehmet Akif Ersoy University Journal of Education Faculty, (45)*, 230-249. <https://dergipark.org.tr/tr/pub/maeuefd/issue/35179/332605>
- De Vellis, R.F. (2003). *Scale development: Theory and applications*. Applied Social Research Methods Series. Sage Publications, Inc. https://www.academia.edu/42875983/Scale_Development_Theory_and_Applications_Second_Edition
- Dogan, E., Guerrero, A., & Tatsuoka, K. (2005). Using DIF to investigate strengths and weaknesses in mathematics achievement profiles of 10 different countries. *In annual meeting of the National Council on Measurement in Education (NCME), Montreal, Canada*. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=a23cbcd509e6d6b9cd664236acc2d585b634578>
- Dinçsoy, L.B. (2022). *Karma testlerde kayıp verilerin değişen madde fonksiyonuna etkisinin incelenmesi [Investigation of the effect of missing data on differential item functioning in mixed type tests]* [Master's dissertation, Hacettepe University]. Hacettepe University. <https://openaccess.hacettepe.edu.tr/xmlui/bitstream/handle/11655/25949/10440993.pdf?sequence=1&isAllowed=y>

- Emenogu, B.C., Falenchuk, O., & Childs, R.A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research*, 56(4), 459- 469. <https://doi.org/10.11575/ajer.v56i4.55429>
- Enders, C.K. (2010). *Applied missing data analysis* (1th ed.). The Guilford Publications, Inc. <http://hsta559s12.pbworks.com/w/file/52112520/enders.applied>
- Erdoğan, K.H. (2019). *Doğrulayıcı faktör analizi ve farklı veri setlerinde uygulanması [Confirmatory factor analysis and application to different datasets]* [Master's dissertation, Applied Sciences University of Isparta]. Higher Education Institution National Thesis Center. https://acikbilim.yok.gov.tr/bitstream/handle/20.500.12812/378756/yokAcikBilim_10284258.pdf?sequence=-1&isAllowed=y
- Garrett, P. (2009). *A Monte Carlo study investigating missing data, differential item functioning, and effect size*. Georgia State University. https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1034&context=eps_diss
- Gök, B., Kabasakal, K.A., & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi [Analysis of attitude items in PISA2009 student questionnaire in terms of differential item functioning based on culture]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 5(1), 72-87. <https://doi.org/10.21031/epod.64124>
- Gültekin, S., & Demirtaşlı, N.Ç. (2020). Comparing the test information obtained through multiple choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online*, 11(1), 251-251. <https://www.ilkogretim-online.org/fulltext/218-1596943363.pdf?1697476130>
- Hambleton, R.K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory (Vol. 2)*. Sage.
- Jabrayilov, R., Emons, W.H., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, 40(8), 559-572. <https://doi.org/10.1177/0146621616664046>
- Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14(4), 329-349. <https://eric.ed.gov/?id=EJ642273>
- Josse, J., Mayer, I., Tierney, N., & Vialaneix, N. (2022). CRAN task view: Missing data. <https://mirror.truenetwork.ru/CRAN/web/views/MissingData.html>
- Kalaycıoğlu, D.B., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Analysis of attitude items in PISA2009 student questionnaire in terms of differential item functioning based on culture]. *Eğitim ve Bilim*, 36(161), 3-13. <http://egitimvebilim.ted.org.tr/index.php/EB/article/view/143/280>
- Kim, S.H., Cohen, A.S., & Kim, H.O. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement*, 18(3), 217-228. <https://doi.org/10.1177/014662169401800303>
- Longford, N.T. (2005). *Missing data and small-area estimation: Modern analytical equipment for the survey statistician*. Springer.
- Magis, D., Beland, S., Raiche, G., & Magis, M.D. (2015). Package 'difR'. <https://cran.r-project.org/web/packages/difR/difR.pdf>
- MEB (2019). *Uluslararası öğrenci değerlendirme programı PISA 2018 ulusal raporu [International student assessment program PISA 2018 national report]*. Ankara: Directorate of Measurement, Evaluation and Testing Services, Ministry of National Education. https://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_pisa_2018_turkiye_on_raporu.pdf
- OECD (2019). *PISA 2018 results volume I: What students know and can do*. OECD Publishing. <https://www.oecd.org/education/pisa-2018-results-volume-i-5f07c754-en.htm>

- Peng, C.Y., Harwell, M.R., Liou, S.M., & Ehman, L.H. (2006). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 31-78).
- Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2), 197-207. <https://conservancy.umn.edu/bitstream/handle/11299/113559/v14n2p197.pdf?sequence=1>
- Rizopoulos, D., & Rizopoulos, M.D. (2018). Package 'ltm'. <https://cran.stat.unipd.it/web/packages/ltm/ltm.pdf>
- Robitzsch, A., & Rupp, A.A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, 69(1), 18-34. <https://doi.org/10.1177/0013164408318756>
- Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17(2), 105-116. <https://doi.org/10.1177/014662169301700201>
- Rosseel, Y., Jorgensen, T.D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (June 17, 2017). Package 'lavaan'. Version 0.6-18. <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf>
- Salaria, N. (2012). Meaning of the term descriptive survey research method. *International Journal of Transformations in Business Management*, 1(6), 1-7.
- Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177. <https://doi.org/10.1037/1082-989X.7.2.147>
- Schafer, J.L., & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 545-571. https://doi.org/10.1207/s15327906mbr3304_5
- 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.
- Selvi, H., & Alici, D. (2018). Investigating the impact of missing data handling methods on the detection of differential item functioning. *International Journal of Assessment Tools in Education*, 5(1), 1-14. <https://files.eric.ed.gov/fulltext/EJ1250131.pdf>
- Sırgancı, G., & Çakan, M. (2020). Sıralı lojistik regresyon ve poly-sıbttest yöntemleri ile değişen madde fonksiyonunun belirlenmesi [Determination of the differential item function with ordered logistic regression and poly-sıbttest methods]. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, 20(1), 705-717. <https://doi.org/10.17240/aibuefd.2020.20.52925-665084>
- Sünbül, S.Ö., & Sünbül, Ö. (2016). Değişen madde fonksiyonunun belirlenmesinde kullanılan yöntemlerde I. Tip hata ve güç çalışması [Type I error rates and power study of several differential item functioning determination methods]. *İlköğretim Online*, 15(3), 882-897. <https://doi.org/10.17051/io.2016.10640>
- Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6. Ed.). Pearson.
- Tamcı, P. (2018). *Kayıp veriyle başa çıkma yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi* [Investigation of the impact of techniques of handling missing data on differential item functioning] [Master's dissertation, Hacettepe University]. Hacettepe University Open Archive. <https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/5315>
- Taş, U.E., Arıcı, Ö., Ozarkan, H.B., & Özgürlük, B. (2016). PISA 2015 ulusal raporu [PISA 2015 national report]. *Ministry of National Education*. https://odsgm.meb.gov.tr/test/analizler/docs/PISA/PISA2015_Ulusal_Rapor.pdf

- Taşkıran, C., & Şenel, E. (2022). Çok boyutlu sportmenlik yönelimi ölçeğinin ölçme eşdeğerliğinin test edilmesi [Testing the measurement invariance of the multidimensional sportspersonship orientation scale]. *International Journal of Sport Exercise and Training Sciences-IJSETS*, 8(4), 190-196. <https://doi.org/10.18826/useeabd.1156699>
- Templ, M., Alfons, A., Kowarik, A., Prantner, B., & Templ, M.M. (2016). VIM: Visualization and Imputation of Missing Values. R package version 4.6.0, URL <https://CRAN.R-project.org/package=VIM>
- Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.
- Uyar, Ş. (2015). *Gözlenen gruplara ve örtük sınıflara göre tanımlananları madde etkilerinin karşılaştırılması [Comparing differential item functioning based on manifest groups and latent classes]* [Doctoral dissertation, Hacettepe University]. Hacettepe University Open Access System. <https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/1816>
- Van de Vijver, F.J., & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, 47(4), 263-279. https://pure.uvt.nl/ws/files/225989/26727_11858.pdf
- Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman & Hall/CRC Press.
- Yılmaz, M. (2021). *Eğilim puanları kullanılarak ABİDE çalışmasındaki maddelerin değişen madde fonksiyonu açısından incelenmesi [Investigation of differential item functioning of the test items in the abide study by using propensity scores]* [Master's dissertation, Hacettepe University]. Hacettepe University Open Access System. <https://openaccess.hacettepe.edu.tr/xmlui/handle/11655/23603>

Adaptation of the quiet quitting scale for teachers to Turkish culture: An empirical psychometric investigation

Müslim Alanoğlu^{1,*}, Songül Karabatak², Alper Uslukaya³, Ayşenur Kuloğlu⁴

¹Firat University, Faculty of Education, Department of Educational Sciences, Elazığ, Türkiye

²Firat University, Faculty of Education, Department of Educational Sciences, Elazığ, Türkiye

³Çankırı Karatekin University, Faculty of Humanities and Social Sciences, Department of Educational Sciences, Çankırı, Türkiye

⁴Firat University, Faculty of Education, Department of Educational Sciences, Elazığ, Türkiye

ARTICLE HISTORY

Received: Feb. 15, 2024

Accepted: May 13, 2024

Keywords:

Quiet quitting,
Teacher,
Scale adaptation.

Abstract: The study aims to introduce to the Turkish culture a measurement tool that has proven validity and reliability in determining the level of quiet quitting among teachers. It involves the analysis of the validity and reliability of the Quiet Quitting Scale, as the scale is adapted to the Turkish culture. The scale, originally developed in English, was adapted to Turkish using data from teachers employed in public schools who were selected through convenience sampling. Confirmatory factor analysis was initially used to assess the construct validity of the original structure of the scale within the Turkish context. The findings indicated a good fit to the four-factor model, supported by adequate factor loadings and fit indices, thus confirming the scale's validity within the Turkish culture. Reliability evaluation included internal consistency coefficients, test-retest stability, and composite reliability, all exceeding the threshold values. The test-retest analysis confirmed the stability of the scale, while the composite reliability analysis further supported its reliability. Measurement invariance across gender and tenure was examined, confirming that the scale can provide reliable comparisons across these demographic groups. Overall, these results demonstrate the successful adaptation of the Quiet Quitting Scale to Turkish culture and are supported by strong evidence of its validity and reliability.

1. INTRODUCTION

It can be argued that individuals are experiencing more negative situations in their professional lives as a result of global disasters, wars, or pandemics, particularly in recent times. These situations can range from job loss to assuming remote work roles or working extensive hours, all of which can result in excessive fatigue, psychological issues, and burnout. To cope with these adversities, employees often develop various defense mechanisms. In the literature, the actions displayed by employees due to burnout resulting from challenging work conditions are referred to as "quiet quitting behavior" (Yıldız & Özmenekşe, 2022).

In its literal sense, "quitting" refers to the voluntary departure or withdrawal from a position (Turkish Language Association, 2024). On the other hand, "quiet quitting" is described as a

*CONTACT: Müslim Alanoğlu ✉ muslimalanoglu@gmail.com 📍 Firat University, Faculty of Education, Department of Educational Sciences, Elazığ, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

disengagement strategy favored by young employees who do not intend to quit but instead choose to reduce their efforts (Duman, 2023). The concept is further explained as simply carrying out assigned tasks within designated working hours (Kont, 2022), whereby employees do only what is necessary for their job and do not devote additional time, effort, or enthusiasm (Daugherty & Kvilhaug, 2022). Generally, quiet quitting involves employees fulfilling their job responsibilities outlined in their job descriptions and declining to go beyond that (Rogers, 2022; Wheeler, 2022).

While quiet quitting is often described as a behavior that has become prominent in recent times, it is noted that it has been a common workplace behavior among employees in previous years (Arar et al., 2023). Initially articulated by economist Mark Blodger at the A&M Economy Symposium in 2009 as a decline in passion for success the phenomenon of quiet quitting gained attention in 2022 through a video shared by TikToker Zaid Khan (Yıkılmaz, 2022). In the video, Khan stated, "Quiet quitting doesn't mean quitting your job. It just means preventing your job from taking over your life. Your job is not your life! Your worth is not defined by what you produce." This explanation garnered significant interest, particularly among Generation Z, drawing more attention to the concept of quiet quitting. Therefore, quiet quitting is expressed as the response of Generations Y and Z, who sacrifice their time, happiness, and health for their jobs (Mamona, 2022; Önder, 2022).

Several factors contribute to quiet quitting, which can be categorized into three main areas: the work environment, managers, and colleagues. Negative attitudes and behaviors exhibited by managers, employee bullying, biased management practices, heavy workload, inadequate compensation, communication problems, neglect and lack of support, inability to cope with workload, feelings of inefficacy, lack of job satisfaction, high expectations, limited personal time, detachment from the work environment, and overall unhappiness have all been identified as potential precursors to quiet quitting (Arar et al., 2023; Chavarin, 2023; Eflatun, 2023).

Quiet quitting, which is contagious, can lead to negative consequences such as decreased productivity, demotivation, and job dissatisfaction if left unchecked (Yıldız, 2023). Both the organizational and individual consequences of quiet quitting make it an important phenomenon that should be highlighted in the literature on organizational management. In the organizational context, quiet quitting can lead to managers pressuring employees, restricting their flexibility, widespread layoffs, the need to seek new personnel, and a disruptive work environment (Cohen, 2022; Güler, 2023; Miller, 2022; Thompson, 2022). At an individual level, quiet quitting can make individuals feel powerless and may result in poor performance and a lack of opportunities to gain experience due to reduced effort. However, quiet quitting can also have some positive consequences for individuals. When the balance between personal and professional lives starts to blur, individuals may resort to quiet quitting to restore this equilibrium. In such cases, quiet quitting can allow employees to take a break and restore balance in their lives (Bansal, 2023). It is also suggested that quiet quitting can be beneficial in terms of preventing burnout, enhancing a sense of control, and helping individuals prioritize what truly matters in life (Kolev, 2022; Scott, 2022).

In this particular context, it is of utmost importance to implement communicative strategies aimed at enhancing communication within the work environment, fostering and consolidating collaboration among employees, disseminating information about career progression, and establishing a sense of shared purpose to mitigate the occurrence of quiet quitting (Elgan, 2022; Hetler, 2022). Moreover, it is imperative to enhance working conditions, cultivate motivational behaviors, ensure equitable rewards, promote workplace flexibility, and cultivate a positive and blissful work environment as additional measures to deter quiet quitting (Güler, 2023). Furthermore, Klotz and Bolino (2022) highlight that incentives such as paid time off, salary increments, employee involvement in decision-making processes, and encouragement of creativity constitute other viable measures to counter quiet quitting.

Quiet quitting behaviors are also observed among teachers in educational institutions. These behaviors can be attributed to changes in organizational and environmental factors, resulting in weakened perceptions of organizational justice, reduced job satisfaction, and burnout. Factors such as increased workloads and high-performance expectations contribute to these outcomes (Yücedađlar et al., 2024). In the education system, particularly in the post-pandemic era, where new skills are in demand, greater attention should be devoted to teachers as a valuable resource. This attention is essential to retain teachers and ensure high levels of efficiency (Morrison-Beedy, 2021). Teachers play a crucial role in facilitating learning, motivating students, and fostering their intellectual and personal growth (Darling-Hammond, 2000). However, the current high expectations placed on teachers generate significant pressures that can lead to emotional exhaustion, decreased motivation, and decreased job satisfaction (Ingersoll & Strong, 2011).

While some educators may choose to leave the profession due to the challenges they face, there is concern regarding those who remain but quietly disengage from their responsibilities. This phenomenon, known as "quiet quitting," is viewed as a form of passive resistance or silent protest by teachers who feel frustrated, unsupported, or overwhelmed (Santoro, 2019). Quiet quitting is characterized by a gradual decline in motivation, enthusiasm, and dedication to teaching. Teachers experiencing this may fulfill their duties without actively engaging with students or performing at their best. This disconnection from the teaching-learning process can significantly impact students' academic achievements, as well as the overall morale and culture within educational institutions (Altun & Vural, 2012).

The concept of quiet quitting has recently emerged as a new phenomenon in organizational behavior. In recent years, there has been increasing interest in the phenomenon of quiet quitting in organizations. As a result, scales have been developed to determine perceptions of quiet quitting among business employees (Boz et al., 2023), local government employees (Avcı, 2023), healthcare workers (Karaşin & Öztirak, 2023), and university students (Savaş & Turan, 2023). However, there is still insufficient explanation regarding its impact on organizations and individuals. Furthermore, there are only a few studies that help us to understand this concept, especially those that focus on teachers. In the Turkish literature, a scale developed by Yücedađlar et al. (2024) has been used to determine the quiet quitting behaviors exhibited by teachers. This scale assesses three sub-dimensions of quiet quitting: job performance, indifference towards school, and desensitization to work. In contrast, a scale developed by Thomas et al. (2022), which has been adapted for the current study, conceptualizes quiet quitting in terms of emotional exhaustion, incentives, work environment, and job satisfaction. The adapted scale aims to explain faculty members' attitudes towards their professions and work environments. By comparing the dimensions of the two scales, it can be concluded that they measure different aspects of the quiet quitting phenomenon. Therefore, the scale developed by Thomas et al. (2022) is distinct from the one developed by Yücedađlar et al. (2024). Furthermore, the presence of different measurement tools is significant in approaching the new phenomenon of quiet quitting from various perspectives. Additionally, adapting an existing scale with established psychometric properties to a new culture is considered safer than developing a new test, which highlights the importance of adaptation studies (Hambleton & Patsula, 1999). Therefore, it is crucial to adapt and conduct further psychometric analyses to assess the validity and reliability of the Quiet Quitting Scale (QQS) developed by Thomas et al. (2022) through a comprehensive study of the Turkish culture. In light of this, the study aims to contribute a valid and reliable measurement tool that can be used to assess public school teachers' attitudes toward quiet quitting in the literature.

2. METHOD

The process of adapting the QQS to Turkish culture included validity and reliability assessments. Initially, confirmatory factor analysis (CFA) was used to validate the scale's

underlying factor structure. The results from this analysis were then supported by both test-retest and parallel test methods, which clarified the reliability measure of the scale. To demonstrate the validity of the scale, the measurement invariance of the QQS was also examined according to gender and tenure categories.

2.1. Research Model

The purpose of this study is to provide evidence of the validity and reliability of the QQS. However, it does not examine any causal relationships. Therefore, it was conducted as a cross-sectional study within the quantitative research paradigm. Cross-sectional studies involve collection of relevant data at one point in time, without considering the passage of time. All data are collected and primarily associated with the time of data collection or a period close to it (Kesmodel, 2018).

2.2. Study Group

This study focused on teachers employed in public schools in Elazığ province, Türkiye during the 2023-2024 academic year. The study group consisted of volunteer teachers working in the Elazığ province. Given the emphasis on scale adaptation, the aim is not to extend the findings to a broader population. Therefore, the convenience sampling method was employed to select participants, ensuring a convenient and efficient process for data collection. A total of 376 teachers were selected to participate in the study. Data for the research was collected at two different points in time. During the initial data collection period (T1), various scales were administered, including the QQS, the Emotional Exhaustion Dimension of the Maslach Burnout Inventory, the Organizational Support Scale, the Perceived Collegial Support Scale, and the Perceived Supervisor Support Scale. After a three-week interval, the QQS was administered again to the same group of 113 individuals and was selected for test-retest reliability (T2). Detailed information about the participants at both time points (T1 and T2) is given in [Table 1](#).

Table 1. Demographic information about participants.

Category	Variables	N	%
T1 (N = 376)			
Gender	Female	209	55.6
	Male	167	44.4
Education level	Bachelor's degree	292	77.7
	Postgraduate	84	22.3
Tenure = 13.52 (sd = 8.99) years			
T2 (N = 113)			
Gender	Female	75	66.4
	Male	38	33.6
Education level	Bachelor's degree	71	62.8
	Postgraduate	42	37.2
Tenure = 10.47 (sd = 7.14) years			

In the first group, 55.6% of the teachers are female ($n = 209$), while 44.4% ($n = 167$) are male; 77.7% ($n = 292$) of the teachers have a bachelor's degree while 22.3% ($n = 84$) have a postgraduate degree. The average tenure of the teachers is 13.52 years (standard deviation = 8.92). In the second group, 66.4% ($n = 75$) of the teachers are female, while 33.6% ($n = 38$) are male; 37.2% ($n = 42$) of the teachers have a postgraduate degree and 62.8% ($n = 71$) have a bachelor's degree with an average tenure of 10.47 years (standard deviation = 7.14).

2.3. Ethical Consideration

The study received ethical approval from the Ethics Committee of Firat University, Social and Humanities Research, on August 3, 2023, with reference number 2023/14. All procedures followed were in accordance with the ethical standards set by the committee, as well as the 1964 Helsinki Declaration and its subsequent revisions (Rickham, 1964).

2.4. Scales and Procedures

The original version of the QQS is in English, was developed by Thomas et al. (2022) for faculty members, and is structured as a five-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). This scale consists of a total of 33 items categorized into four sub-scales: *emotional exhaustion* (Cronbach's alpha =.92), *incentives* (Cronbach's alpha =.933), *work environment* (Cronbach's alpha =.955), and *job satisfaction* (Cronbach's alpha =.901). The adaptation process followed the recommended procedures outlined in the literature, including needs assessment, selection of an appropriate scale, translation into the target language, back-translation, initial linguistic validation, administration to the study group, validation, reliability analyses, and reporting (Hambleton & Patsula, 1999; International Test Commission, 2017; Seğer, 2015).

Permission was obtained from the scale developer to adapt the scale into Turkish using the back-translation method (Brislin, 1970). The translation of scale items into Turkish was carried out by researchers and reviewed by four faculty members, consisting of two experts in the Educational Administration Department and two in the Educational Measurement and Evaluation Department. Following their feedback, the revised items were scrutinized by two Turkish language experts. Subsequently, the translated items were back-translated into English and a comparison with the original scale was conducted by two English language experts to ensure fidelity of meaning. Necessary adjustments were made based on their recommendations. A pilot study was then conducted with 30 teachers to assess the clarity of the items, leading to the finalization of the Turkish version of the scale for implementation (see the Turkish version of the QQS in the [Appendix](#)).

To ensure the nomological validity of the scale, parallel scales that are theoretically associated with the QQS and its sub-scales were utilized. To assess the initial subscale of the QQS, the nine items of the Burnout Scale, originally formulated by Maslach and Jackson (1981) and later adapted into Turkish by Ergin (1992), were employed as a parallel test. The second sub-scale, incentives, consists of items of the support that teachers receive in their roles. Accordingly, the short form of the Organizational Support Scale, developed by Eisenberger et al. (1986) and comprising eight items, was used as a parallel test for this sub-scale. The third sub-scale, work environment, was assessed using the Perceived Collegial Support Scale, developed by Oranje (2001) and adapted into Turkish by Özgün (2005). This parallel test comprised six items. Lastly, the fourth sub-scale, job satisfaction, was evaluated using the Perceived Supervisor Support Scale, developed by Magill (2002) and adapted into Turkish by Özgün (2005), which included seven items.

2.5. Data Analysis

Analyses were conducted using SPSS 27 and Mplus version 8.10. First, the data collected was examined for any missing values. Subsequently, the values of kurtosis and skewness were assessed. However, the results of the test for multivariate normality demonstrated that the Mardia's skewness (174.31; $p = .00$) and kurtosis (1334.21; $p = .00$) values were statistically significant, indicating a failure to meet the assumption of multivariate normality. Consequently, the maximum likelihood estimator with robust standard errors (MLR) method was employed as the parameter estimation approach in CFA (Muthén & Muthén, 1998-2017; Şen, 2023). Subsequently, the mean and standard deviation values of the scale/dimension structures of the

data were computed. To reveal the suitability of the scale for Turkish culture, analyses on the validity and reliability of the scale structure were conducted.

A CFA was conducted to examine the four-factor structure of the QQS. The fit criteria used to assess model fit in CFA included the chi-square/degree of freedom (χ^2/df), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error (RMSEA), and Standardized Root Mean Squared Residual (SRMR) (Xu & Tracey, 2017). To indicate a good fit in CFA, the χ^2/df ratio should be less than 3, CFI and TLI values should be greater than .90, and RMSEA and SRMR values should be less than 0.08 (Hu & Bentler, 1999). These compliance criteria were taken into account in the CFA sections.

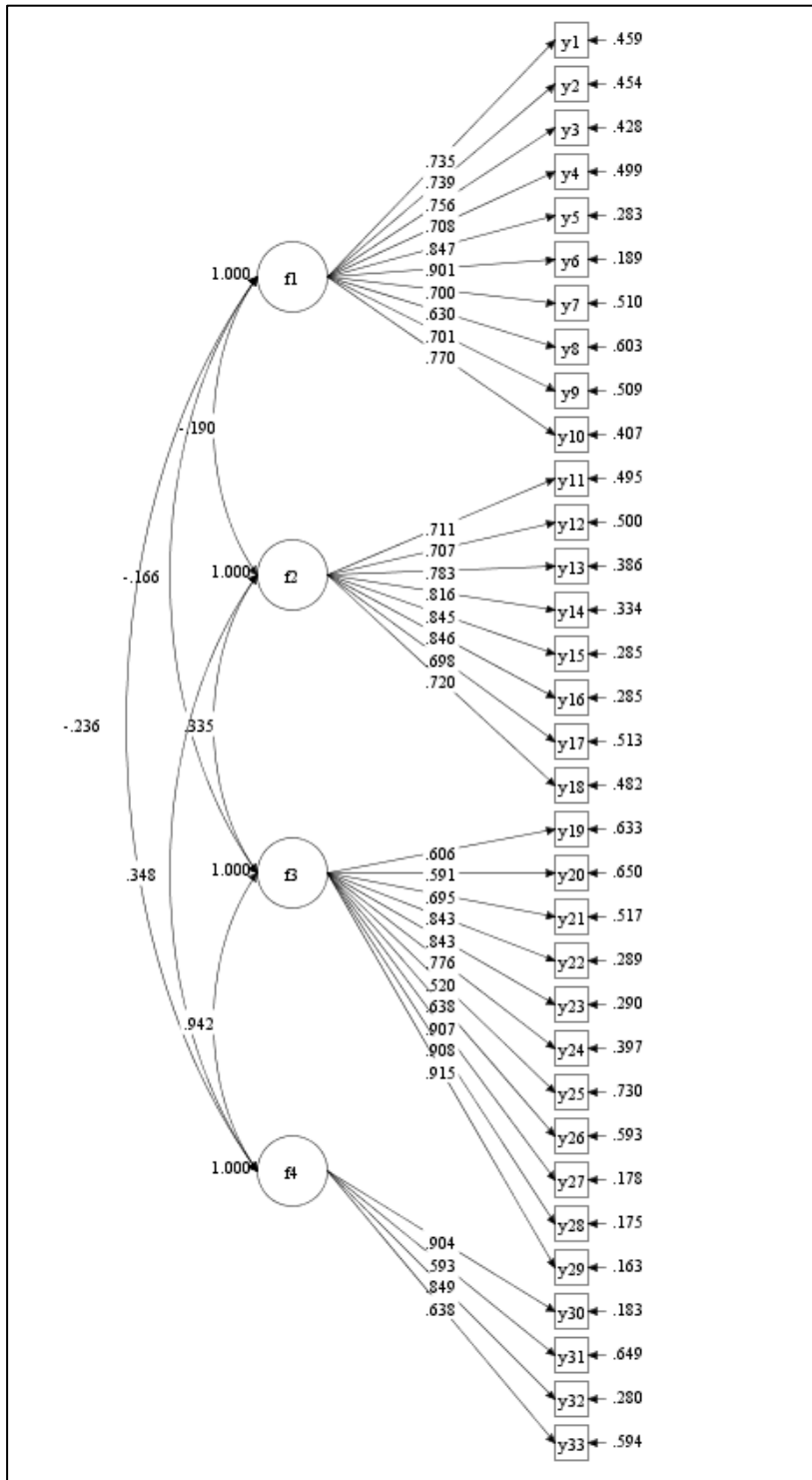
To assess the internal reliability of the scale, Cronbach's Alpha and McDonald's Omega coefficients were calculated. A value of .70 or higher for these coefficients was considered acceptable for internal consistency (Hayes & Coutts, 2020; McDonald, 2013). To support these values, composite reliability (CR) and average extracted variance (AVE) were calculated based on CFA factor loadings. Test-retest values were evaluated to examine the stability of the scale, CR and AVE values were evaluated to determine convergent validity, and parallel test values were evaluated for nomological validity. Test-retest reliability was ensured by maintaining stable significant results at the $p < 0.01$ level (Gravesande et al., 2019). A correlation value of .50 or higher was accepted in parallel tests (Cohen, 1988). Additionally, the fact that CR values were higher than AVE values and that the $AVE > .50$ served as evidence of convergent validity (Fornell & Larcker, 1981). Discriminant validity is achieved when the square root of the AVE is greater than the correlation between constructs (Zainudin, 2012).

Measurement invariance was assessed simultaneously for the QQS and the estimated CFA model. Typically, measurement invariance is determined by examining the change in χ^2 (Byrne et al., 1989). Muthén and Muthén (2012) suggest that non-significant results should be evaluated for greater parsimony compared to the more constrained model, which assumes a certain level of stability but fits equally well. However, it is important to note that the size of the intervals affects the χ^2 values, and thus a "perfect" model is highly sensitive to intermittent errors, particularly over large areas (Chen, 2007). Consequently, the presence of various fit indices becomes crucial when comparing the two nested models. Cheung and Rensvold (2002) indicate that a change of -.01 in CFI can be considered to ensure measurement invariance; however, it is also suggested that alternative fit indices such as $\Delta RMSEA$ and $\Delta SRMR$ can be used to evaluate the measurement stability of certain components (Meade et al., 2008). Chen (2007) found that ΔCFI and ΔTLI should be at least .01, whereas he recommends utilizing .015 as the threshold for $\Delta RMSEA$ and $\Delta SRMR$. In the current study, we aimed to determine whether the quiet quitting behavior exhibits measurement invariance across gender (males vs. females) and tenure (below 13 years vs. 13 and above years) categories. Given that the average tenure of the participants was 13.52 and there were approximately an equal number of participants with tenure below and above this value, the participants were divided into two groups: those with tenure below 13 years ($n = 194$) and those with tenure above 13 years ($n = 182$). To achieve invariance, we assessed the ΔCFI , ΔTLI , $\Delta RMSEA$, and $\Delta SRMR$ criteria in addition to the chi-square difference test.

3. FINDINGS

In this section, we present the results obtained from the scale validity and reliability, as well as measurement invariance, consecutively. The DFA diagram related to the four-factor structure of the QQS is presented in [Figure 1](#).

Figure 1. CFA model for the QQS.



The results of the DFA model indicate that the scale effectively was adapted to Turkish culture, confirming the four-factor structure of the QQS. The robust indices obtained provide support for this conclusion, including $\chi^2 = 1223.761$ ($df = 489$; $p = .000$), RMSEA = 0.063 (90% CIs = 0.059-0.068), CFI = .917, TLI = .911, and SRMR = 0.047. Table 2 presents a comprehensive overview of the DFA results, including the values for Cronbach's alpha, McDonald's omega, CR, and AVE.

Table 2. CFA results and reliability values of the QQS.

Sub-Scales	Item No	QQ1	QQ2	QQ3	QQ4	S.E.	<i>z</i>	<i>p</i>	Cronbach alpha	McDonald's omega	CR	AVE
QQ1	Item1	.735				0.026	28.733	.000	.927	.927	.928	.566
	Item2	.739				0.025	29.173	.000				
	Item3	.756				0.024	31.479	.000				
	Item4	.708				0.028	25.673	.000				
	Item5	.847				0.017	50.296	.000				
	Item6	.901				0.012	72.833	.000				
	Item7	.700				0.028	24.754	.000				
	Item8	.630				0.033	19.099	.000				
	Item9	.701				0.028	24.816	.000				
	Item10	.770				0.023	33.542	.000				
QQ2	Item11		.711			0.028	25.451	.000	.918	.918	.920	.590
	Item12		.707			0.028	24.973	.000				
	Item13		.783			0.023	34.680	.000				
	Item14		.816			0.020	41.007	.000				
	Item15		.845			0.018	47.867	.000				
	Item16		.846			0.018	47.812	.000				
	Item17		.698			0.029	24.262	.000				
	Item18		.720			0.027	26.533	.000				
QQ3	Item19			.606		0.034	17.977	.000	.938	.940	.923	.555
	Item20			.591		0.035	17.078	.000				
	Item21			.695		0.028	25.049	.000				
	Item22			.843		0.016	52.522	.000				
	Item23			.843		0.016	52.361	.000				
	Item24			.776		0.022	36.028	.000				
	Item25			.520		0.039	13.449	.000				
	Item26			.638		0.032	20.168	.000				
	Item27			.907		0.010	87.158	.000				
	Item28			.908		0.010	87.944	.000				
	Item29			.915		0.010	94.015	.000				
QQ4	Item30				.904	0.013	71.843	.000	.850	.851	.839	.574
	Item31				.593	0.036	16.660	.000				
	Item32				.849	0.017	49.229	.000				
	Item33				.638	0.033	19.234	.000				
QQS									.877	.783	.977	.576

Note(s): QQ1. Emotional Exhaustion; QQ2. Incentives; QQ3. Work Environment; QQ4. Job Satisfaction; QQS. Quiet Quitting Scale

The factor loadings of the CFA model presented in Table 2 range from .520 to .908. Furthermore, all standard loadings of the factors demonstrate statistical significance, with *z*-values exceeding 2.56 and *p*-values less than .01. The reliability of the sub-scales is evaluated using Cronbach's alpha, McDonald's omega, and organic reliability values, which serve as the required threshold values. Table 3 shows the mean and standard deviation values of the scales used throughout the study, as well as the findings regarding the validity and reliability of the QQS scale.

Table 3. Validity and reliability analysis results for the QQS.

T1. Parallel test (N = 376)							Discriminant validity (N = 376)				T2. Test-retest (N = 113)			
	Mean	SD	QQ1	QQ2	QQ3	QQ4	QQ1	QQ2	QQ3	QQ4	QQ1	QQ2	QQ3	QQ4
QQ1	2.88	1.025	-				.75				.66**			
QQ2	2.84	0.913	-.16**	-				.77				.58**		
QQ3	3.55	0.902	-.26**	.29**	-				.74				.56**	
QQ4	3.49	0.950	-.23**	.34**	.81**	-				.76				.51**
EE	2.43	0.979	.66**	-.20**	-.23**	-.25**								
OS	3.46	0.914	-.30**	.51**	.70**	.61**								
PCS	3.04	0.592	-.14**	.07	.55**	.48**								
PSS	2.96	0.837	-.24**	.23**	.71**	.64**								

** $p < .01$; QQ1. Emotional Exhaustion; QQ2. Incentives; QQ3. Work Environment; QQ4. Job Satisfaction; EE. Maslach Burnout Inventory Emotional Exhaustion; OS. Organizational Support; PCS. Perceived Colleague Relations Support; PSS. Perception of Supervisor Support.

The study reveals significant relationships between different scales. Firstly, the emotional exhaustion subscale of the QQS demonstrates a positive correlation with the emotional exhaustion dimension of the Maslach Burnout Inventory ($r = .66$; $p < .01$), indicating a moderate association. Secondly, the incentives subscale of the QQS is positively correlated with the Organizational Support Scale ($r = .51$; $p < .01$). Additionally, the Perceived Colleague Relations Support Scale shows a positive correlation with the work environment subscale ($r = .55$; $p < .01$), indicating a notable relationship. Finally, the Job Satisfaction subscale is positively correlated with the Perceived Supervisor Support Scale ($r = .64$; $p < .01$), demonstrating a significant association. These findings emphasize the convergent validity of the QQS, as its correlation values exceed the accepted threshold of $r = .50$ ($p < .01$). Moreover, the results ensure the nomological validity, and QQS achieves convergent validity through the CR/AVE values. The square root of the AVE values showed that discriminant validity was achieved.

When examining the test-retest correlation values among the sub-scales of the QQS, we observed that there were correlation values ($r > .50$; $p < .01$) for emotional exhaustion ($r = .66$; $p < .01$), incentives ($r = .58$; $p < .01$), work environment ($r = .56$; $p < .01$), and job satisfaction ($r = .51$; $p < .01$) sub-scales. The test-retest reliability of the QQS was found to be sufficient. The categories determined by gender and tenure variables were evaluated in terms of the four levels of measurement invariance; namely, configural, metric, scalar, and strict. The results are presented in Table 4. *Tests for gender invariance* yielded the following fit statistics for the different models: the configural model had $\chi^2(978) = 1766.493$, CFI = .913, TLI = .906, RMSEA = 0.066, and SRMR = 0.060. For the metric model, the values were $\chi^2(1007) = 1803.961$, CFI = .912, TLI = .907, RMSEA = 0.065, and SRMR = 0.064, indicating invariance. Similarly, the scalar model showed $\chi^2(1036) = 1841.126$, CFI = .911, TLI = .909, RMSEA = 0.064, and SRMR = 0.066, confirming invariance. Lastly, the strict model displayed $\chi^2(1069) = 1893.127$, CFI = .909, TLI = .910, RMSEA = 0.064, and SRMR = 0.066, confirming invariance. Therefore, the dataset met the requirement for invariance of the gender measure across the metric, scalar, and strict models. This is supported by insignificant χ^2 difference tests and consistent changes in CFI, TLI, RMSEA, and SRMR.

As for tenure invariance, the fit indices for the configural model were $\chi^2(978) = 1842.040$, CFI = .905, TLI = .897, RMSEA = 0.068, and SRMR = 0.062. For the metric model, the values were $\chi^2(1007) = 1880.638$, CFI = .904, TLI = .899, RMSEA = 0.068, and SRMR = 0.062, indicating invariance. Similarly, the scalar model had $\chi^2(1036) = 1905.958$, CFI = .904, TLI = .902, RMSEA = 0.067, and SRMR = 0.062, indicating invariance. Lastly, the strict model exhibited $\chi^2(1069) = 1951.503$, CFI = .903, TLI = .904, RMSEA = 0.066, and SRMR = 0.063, confirming invariance. Thus, the dataset met the requirement for invariance of the tenure measurement across the metric, scalar, and strict models. This is supported by insignificant χ^2 difference tests and consistent changes in CFI, TLI, RMSEA, and SRMR.

Table 4. Measurement model results.

Model	$\chi^2(df)$	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2(df)$	$p(\chi^2)$	Δ CFI	Δ TLI	Δ RMSEA	Δ SRMR
Gender (N = 376)											
Model 1: Full Configural	1766.493(978)	.913	.906	.066	.060	-	-	-	-	-	-
Model 2: Full Metric	1803.961 (1007)	.912	.907	.065	.064	37.468(29)	.135	-.001	.001	-.001	.004
Model 3: Full Scalar	1841.126 (1036)	.911	.909	.064	.066	37.165(29)	.142	-.001	.002	-.001	.002
Model 4: Full Strict	1893.127(1069)	.909	.910	.064	.066	52.001(33)	.139	-.002	.001	.000	.000
Tenure (N = 376)											
Model 1: Full Configural	1842.040(978)	.905	.897	.068	.062	-	-	-	-	-	-
Model 2: Full Metric	1880.638(1007)	.904	.899	.068	.062	38.598(29)	.110	-.002	.004	-.002	.000
Model 3: Full Scalar	1905.958(1036)	.904	.902	.067	.062	25.319(29)	.662	-.01	-.005	.002	.003
Model 4: Full Strict	1951.503(1069)	.903	.904	.066	.063	45.545(33)	.239	-.002	.002	-.001	.001

4. DISCUSSION and CONCLUSION

This study aims to adapt the QQS developed by Thomas et al. (2022) to Turkish culture and to evaluate the validity and reliability of this adaptation by integrating an international measurement tool into a local context. The original scale, developed in English, was designed to determine faculty members' quiet quitting attitudes; however, in this adaptation study, the analyses were conducted using teachers' data. The validity and reliability analyses of the scale were conducted with a multi-perspective approach. First, a CFA was performed to determine the construct validity of the original structure of the scale in the Turkish culture. The nomological validity of the scale was determined by the parallel test method. Then, the CR and AVE values were evaluated together to determine the convergent validity. Regarding the scale's reliability, stability was tested using the test-retest method, internal consistency was tested using Cronbach's alpha and McDonald's omega coefficients, and composite reliability was tested. Finally, the measurement invariance of the scale was examined based on gender and tenure variables.

The fit indices for the CFA of the scale (χ^2/df , RMSEA, CFI, TLI, and SRMR) indicate a good fit to the four-factor measurement model of the scale. Furthermore, the z -values for the factor loadings of the scale items also demonstrate that all factor loadings are significant. This finding is interpreted as evidence that the construct validity of the scale is established in Turkish culture. The statistically significant factor loadings for each dimension of the QQS can also be considered as evidence of convergent validity (O'Rourke & Hatcher, 2013). In line with this, the factor loadings of the measurement model estimated by the CFA, along with the computed CR and AVE values, provide further evidence that the scale meets the conditions for convergent validity. The scales/dimensions applied for nomological validity, under the expectation that they represent theoretically similar constructs to the sub-scales of the QQS, confirm this expectation and demonstrate that these constructs are empirically related, thus indicating the nomological validity of the scale.

In terms of reliability, the measured internal consistency coefficients (Cronbach's α and McDonald's ω) are above the threshold value for each dimension and the coefficients are close to each other, indicating that the scale is reliable (Kline, 2015). Moreover, the CR values above the threshold value for composite reliability are considered as evidence of the scale's composite reliability. Finally, the significant correlation values among the sub-scales, measured with a three-week interval to test their stability, indicate that the scale is a reliable measure of stability.

The QQS was also evaluated from the perspective of measurement invariance between intervals separated by two variables, such as gender and tenure. It is important to determine whether this assessment measures the same construct across productive groups (Millsap, 2011). Because measurement invariance, such as measurement or sub-measurement averages, can be meaningfully compared between different groups, appropriate measurement stability can be achievable. For both *gender* and *tenure* variables, measurement invariance is met up to the level of full strict invariance. This indicates that the differences in the means observed in the quiet quitting responses between the groups of gender and tenure variables reflect differences in the latent factors measuring teachers' attitudes towards quiet quitting (emotional exhaustion, incentives, working environment, and job satisfaction). For effective modification of group factors, it is imperative to adhere to strict stability conditions. When evaluating differences in latent factor means, the differences in intercepts exhibit the most significant level of performance (Chen, 2007, 2008; Schmitt & Kuljanin, 2008). The results of measurement invariance are important for demonstrating the reliability of the outcomes of differential tests conducted based on gender and tenure variables using the QQS.

4.1. Limitations

This study has limitations and offers valuable insights for future research. The current study did not examine the temporal invariance of the QQS (longitudinal measurement invariance). Since individuals' attitudes towards quiet quitting may change over time, it would be valuable to update the measurement of this construct by capturing changes in behavior and attitudes throughout the process. In other words, items that contribute to muting in modifiers and wide spacing among individuals should be revised (Chen, 2008). Therefore, as an extension of the current study, it would be worthwhile to investigate the longitudinal invariance of the scale to evaluate changes in performance over time (Millsap & Cham, 2013). Another limitation of this study is its limited geographical scope, as it was conducted in only one province. By expanding the study to include teachers from various cities, the generalizability of the findings of the study can be ensured.

4.2. Conclusion

The results of this study indicate that the QQS is a reliable and valid tool for evaluating attitudes towards quiet quitting and shows potential for future development in the Turkish context. While high scores indicating emotional exhaustion suggest a high level of quiet quitting, low scores in the dimensions of incentives, working environment, and job satisfaction also suggest a high level of quiet quitting. The quiet quitting scale, with its potential to quantify the quiet quitting attitudes of teachers, holds significant importance in furthering our understanding of this emerging phenomenon in organizational behavior. With the help of this scale, individuals can offer insight into quiet quitting that may occur due to unfavorable processes within Türkiye. In addition, the scale can help policymakers and educational administrators to understand and take measures to address the phenomenon of quiet quitting, which is likely to lead to negative consequences such as teacher inefficiency and low performance.

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:** Firat University, 2023-14.

Contribution of Authors

Muslim Alanođlu: Software, Formal Analysis, Writing-original Draft, Supervision, Validation. **Songül Karabatak:** Collected Data, Investigation, Resources, Visualization, Software, Formal Analysis. **Alper Uslukaya:** Collected Data, Methodology, and Validation. **Ayşenur Kulođlu:** Collected Data, Methodology, and Validation.

Orcid

Muslim Alanođlu  <https://orcid.org/0000-0003-1828-4593>

Songül Karabatak  <https://orcid.org/0000-0002-1303-2429>

Alper Uslukaya  <https://orcid.org/0000-0003-1455-8438>

Ayşenur Kulođlu  <https://orcid.org/0000-0003-0217-8497>

REFERENCES

- Altun, T., & Vural, S. (2012). Evaluation of the views of teachers and administrators of a science and art center (SAC) about professional development and school improvement. *Electronic Journal of Social Sciences*, 11(42), 152-177.
- Arar, T., Çetiner, N., & Yurdakul, G. (2023). Quiet quitting: Building a comprehensive theoretical framework. *Journal of Academic Researches and Studies*, 15(28), 122-138.
- Avcı, N. (2023). The relations between organizational cynicism, organizational silence, presenteeism and quiet quitting: The case of Istanbul Maltepe Municipality. *Süleyman Demirel University Visionary Journal*, 14(39), 968-989.

- Bansal, V. (2023). *Is quiet quitting a good idea?* <https://www.techtello.com/quiet-quitting/>
- Boz, D., Duran, C., Karayaman, S., & Deniz, A. (2023). *Sessiz istifa ölçeği [Quiet quitting scale]*. C. Duran (Ed.), In *Sessiz istifa [Quiet quitting]* (pp. 10-41). Eğitim yaymevi.
- Brislin, R.W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1, 185–216.
- Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466. <https://doi.org/10.1037/0033-2909.105.3.456>
- Chavarin, G. (2023). *Re-framing quiet quitting: An opportunity for employers to improve the workplace.* <https://www.modernhealth.com/post/what-is-quiet-quitting>
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464-504. <https://doi.org/10.1080/10705510701301834>
- Chen F.F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, 95(5), 1005-1015. <https://doi.org/10.1037/a0013193>
- Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, 9(2), 233-255.
- Cohen, J. (2022). *Quiet quitting: the newest way to strike back against corporate greed.* TMH The Miami Hurricane. <https://www.themiamihurricane.com/2022/09/21/quiet-quitting-the-newest-way-to-strike-back-against-corporate-greed/>
- Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education*, 51, 166 - 173. <https://doi.org/10.1177/0022487100051003002>
- Daugherty, G., & Kvilhaug, S. (2022). What is quiet quitting-and is it a real trend? <https://www.investopedia.com/what-is-quiet-quitting-6743910>
- Duman, C. (2023). Büyük istifadan sonra sessiz istifa. Independent Türkçe. <https://www.indyurk.com/node/547061/t%C3%BCrki%C3%87yedensesler/b%C3%BCy%C3%BCk-istifadan-sonra-sessiz-istifa>
- Eflatun, M. (2023). Quiet quitting as a new concept: Characteristics, recognition and prevention. *Journal of Academic Analysis*, 1(1), 17-31.
- Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology*, 71(3), 500-507.
- Elgan, M. (2022). *It's time to quit quitting on the quiet quitters.* <https://www.computerworld.com/article/1613093/its-time-to-quit-quitting-on-the-quiet-quitters.html>
- Ergin, C. (1992). *The doctors and nurses in burnout and adaptation of the Maslach burnout scale.* VII. National Psychology Congress scientific studies, 22-25 September, Ankara.
- Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50.
- Gravesande, J., Richardson, J., Griffith, L., & Scott, F. (2019). Test-retest reliability, internal consistency, construct validity and factor structure of a falls risk perception questionnaire in older adults with type 2 diabetes mellitus: A prospective cohort study. *Archives of physiotherapy*, 9(1), 1-11.
- Güler, M. (2023). A new concept in work culture: Quiet quitting. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 32(1), 247-261. <https://doi.org/10.35379/cusosbil.1200345>
- Ingersoll, R.M., & Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Educational Research*, 81(2), 201-233.
- Hambleton, R.K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, 1, 2. (online). <http://www.testpublishers.org/journal.html>

- 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.
- Hetler, A. (2022). *Quiet quitting explained: Everything you need to know*. <https://www.techtarget.com/whatis/feature/Quiet-quitting-explained-everything-you-need-to-know>
- Hu, L.-t., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55. <https://doi.org/10.1080/10705519909540118>
- International Test Commission. (2017). *The ITC Guidelines for Translating and Adapting Tests* (Second edition). www.InTestCom.org
- Karaşin, Y., & Öztırak, M. (2023). Development of quiet quitting attitude scale. *Cankırı Karatekin University Journal of the Faculty of Economics and Administrative Sciences*, 13(4), 1443-1460. <https://doi.org/10.18074/ckuiibfd.1311522>
- Kesmodel, U.S. (2018). Cross-sectional studies—what are they good for?. *Acta obstetricia et gynecologica Scandinavica*, 97(4), 388-393.
- Kline, R.B. (2015). *Principles and practice of structural equation modeling*. The Guilford Press.
- Klotz, A.C. & Bolino, M.C. (2022). *When quiet quitting is worse than the real thing*. Harvard business review. <https://hbr.org/2022/09/when-quiet-quitting-is-worse-than-the-real-thing>
- Kolev, G. (2022). *What is “quiet quitting” (and should you join the trend)*. Officetopics.com. <https://officetopics.com/what-is-quiet-quitting/>
- Kont, B. (2022). *What is quiet quitting? Reasons and solutions for companies*. <https://www.ixtalent.com/comprehensive-approach-what-does-quiet-quitting-mean-to-companies/>
- Magill, A.L. (2002). *Studying the needs and experiences of beginning teachers* [Unpublished master's thesis]. University of Alberta.
- Mamona, S. (2022). *'Quiet quitting' is TikTok's antidote to generation burnout – but it only works for the privileged*. <https://www.glamourmagazine.co.uk/article/quiet-quitting-is-a-privilege>
- Maslach, C. & Jackson, S.E. (1981). *MBI Maslach burnout inventory ('Human services survey') Research Edition, Manual*. Consulting Psychologist Press Inc.
- McDonald, R.P. (2013). *Test theory: A unified treatment*. Psychology press.
- Meade, A.W., Johnson, E.C., & Braddy, P.W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, 93, 568-592. <https://doi.org/10.1037/0021-9010.93.3.568>
- Miller, K.L. (2022). *Actually, we've been 'quiet quitting' and 'quiet firing' for years*. The Washington Post. <https://www.washingtonpost.com/business/2022/09/08/quiet-quitting-quiet-firing-what-to-do/>
- Millsap, R.E. (2011). *Statistical approaches to measurement invariance*. Routledge.
- Millsap, R.E., & Cham, H. (2013). Investigating factorial invariance in longitudinal data. In B. Laursen, T.D. Little, & N.A. Card (Eds.), *Handbook of developmental research methods* (pp. 109-148). Guilford Press.
- Morrison-Beedy, D. (2021). Building healthy academic communities in this post-pandemic brave new world. *Building Healthy Academic Communities Journal*, 5(2), 7. <https://doi.org/10.18061/bhac.v5i2.8700>
- Muthén, L.K., & Muthén, B.O. (2012). Latent variable modeling using Mplus: Day 2. <https://www.statmodel.com/download/handouts/Beijing2012-Day2.pdf>
- Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide* (Eighth Edition). Muthén & Muthén.
- Oranje, A.H. (2001). Teacher shortages, teacher job satisfaction, and professionalism: Teacher assistants in Dutch secondary schools. *Reports Research*, (143). <https://files.eric.ed.gov/fulltext/ED453194.pdf>

- O'Rourke, N., & Hatcher, L. (2013). *A step-by-step approach to using SAS for factor analysis and structural equation modeling* (2nd ed.). SAS Institute Inc.
- Önder, N. (2022). Herkesin konuđuđu sessiz istifa nedir? [What is the silent resignation that everyone is talking about?] <https://www.marketingturkiye.com.tr/haberler/sessiz-istifa-nedir>
- Özgün, Ö. (2005). *The relationship of novice Turkish early childhood education teachers? professional needs, experiences, efficacy beliefs, school climate for promoting early childhood learning, and job satisfaction* [Unpublished PhD Thesis]. Syracuse University.
- Rickham P.P. (1964). Human experimentation. Code of ethics of the world medical association (Declaration of Helsinki). *Can Med Assoc J*, 2(5402), 177.
- Rogers, G. (2022). When will "quiet quitting" impact student success? <https://www.edsights.io/post/when-will-quiet-quitting-impact-student-success>
- Santoro, D. (2019). The problem with stories about teacher "burnout". *Phi Delta Kappan*, 101, 26-33. <https://doi.org/10.1177/0031721719892971>
- Savaş, B.Ç., & Turan, M. (2023). Quiet quitting scale: Validity and reliability study. *The Online Journal of Recreation and Sports*, 12(3), 442-453.
- Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18(4), 210-222.
- Scott, E. (2022). Could 'quiet quitting' your job be the answer to burnout? What you need to know. <https://metro.co.uk/2022/07/29/could-the-quiet-quitting-trend-be-the-answer-to-burnout-what-you-need-to-know-17085827/>
- Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci: SPSS ve LISREL uygulamaları* [Psychological test development and adaptation process: SPSS and LISREL applications]. Anı Publication.
- Şen, S. (2020). *Mplus ile yapısal eşitlik modellemesi uygulamaları* [Structural equation modeling applications with Mplus]. Nobel Academy.
- Thomas, A., Prater, E., Jones, E., Whiteside, M., & Reyes, P. (2022). A research view of quiet quitting for gig faculty in academic faculty positions: Part 2 of 3. *International Supply Chain Technology Journal*, 8(2). <https://doi.org/10.20545/isctj.v08.i02.01>
- Thompson, D. (2022). Quiet quitting is a fake trend *The Atlantic*. <https://www.theatlantic.com/newsletters/archive/2022/09/quiet-quitting-trend-employee-disengagement/671436/>
- Türk Dil Kurumu. (2024). Türk Dil Kurumu güncel sözlük [Turkish Language Association current dictionary]. <https://sozluk.gov.tr/>
- Wheeler, M.A. (2022). Quiet quitting: A path to work engagement? *Psychology today*. <https://www.psychologytoday.com/us/blog/ethically-speaking/202209/quiet-quitting-path-work-engagement>
- Xu, H., & Tracey, T.J. (2017). Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. *The European Journal of Counselling Psychology*, 6(1), 75-82. <https://doi.org/10.5964/ejcop.v6i1.120>
- Yıkılmaz, İ. (2022). Quiet quitting: A conceptual investigation. *Anadolu 10th International Conference on Social Sciences*, 15-16 Ekim, Diyarbakır, Türkiye.
- Yıldız, S. (2023). Quiet quitting: Causes, consequences, and suggestions. *International Social Mentality and Research Thinkers Journal*, 9(70), 3180-3190.
- Yıldız, S., & Özmenekşe, Y.O. (2022). An aftermath for worklife: Quiet quitting. *Al-Farabi International Journal on Social Sciences*, 7(4), 14-24.
- Yücedağlar, A., Gılıç, F., Uzun, N.B., & İnanđı, Y. (2024). The quiet quitting of teachers: A validity and reliability study. *Mehmet Akif Ersoy University Journal of Education Faculty*, 69, 227-251.
- Zainudin, A. (2012). *A handbook on SEM: Structural equation modelling using AMOS graphics*. University Technology MARA Press.

APPENDIX: Teacher Quiet Quitting Scale - Turkish version

Madde No	Öğretmen Sessiz İstifa Ölçeđi	Kesinlikle Katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle Katılıyorum
1.	Öğretme sorumluluklarımdan dolayı kendimi duygusal olarak yıpranmış hissediyorum	1	2	3	4	5
2.	Görev ve sorumluluklarımdan dolayı kendimi duygusal olarak yıpranmış hissediyorum	1	2	3	4	5
3.	İş gününün sonunda kendimi tükenmiş hissediyorum	1	2	3	4	5
4.	Sabah kalkıp yeni bir iş günüyle yüzleşmek zorunda kaldığımda kendimi yorgun hissediyorum	1	2	3	4	5
5.	Öğretme sorumluluklarımdan dolayı kendimi tükenmiş hissediyorum	1	2	3	4	5
6.	Görev sorumluluklarımdan dolayı kendimi tükenmiş hissediyorum	1	2	3	4	5
7.	Ders yükümün fazla olduğunu düşünüyorum	1	2	3	4	5
8.	Sorumlu olduğum dersler için çok fazla çalıştığımı düşünüyorum	1	2	3	4	5
9.	Görevlerimden dolayı çok fazla çalıştığımı düşünüyorum	1	2	3	4	5
10.	Dayanma gücümün son noktasındaymışım gibi hissediyorum	1	2	3	4	5
11.	Araştırmalarımla hakkında oldukça fazla geri bildirim alırım	1	2	3	4	5
12.	Öğretme becerilerimle ilgili önemli ölçüde geri bildirim alırım	1	2	3	4	5
13.	Okula verdiğim hizmet hakkında oldukça fazla geri bildirim alırım	1	2	3	4	5
14.	Araştırmalarımla kalitesi konusunda önemli ölçüde destek alırım	1	2	3	4	5
15.	Öğretim faaliyetlerimin kalitesiyle ilgili oldukça fazla miktarda rehberlik sağlanır	1	2	3	4	5
16.	Yerine getirdiğim hizmetlerin kalitesiyle ilgili bana büyük ölçüde kılavuzluk edilir	1	2	3	4	5
17.	Öğretmenlik mesleğinin maddi olarak tatmin edici olduğunu düşünüyorum	1	2	3	4	5
18.	Okulumun sağladığı avantajlar yaptığım çalışmalardan daha büyük etkiye sahiptir	1	2	3	4	5
19.	Okulumda yönetici ve öğretmenler arkadaş canlısıdır	1	2	3	4	5
20.	Okulumda arkadaş edinmem için bana fırsatlar verilir	1	2	3	4	5
21.	Okulumda kişisel olarak önemsendiğimi hissediyorum.	1	2	3	4	5
22.	Okul yönetimi, öğretmenlerin birlikte çalışmalarını sağlama konusunda başarılıdır	1	2	3	4	5
23.	Okul yönetimi, öğretmenlik hizmetini yerine getirmemde bana yardımcı olmaktadır	1	2	3	4	5
24.	Okul yöneticiler, yerine getirmem gereken görevlerde bana yardımcı olmaktadır	1	2	3	4	5
25.	Okulumdaki öğretmenler araştırmalarımda bana yardımcı olmaktadır	1	2	3	4	5
26.	Okulumdaki öğretmenler yerine getirmem gereken görevlerde bana yardımcı olmaktadır	1	2	3	4	5
27.	Okul yönetimi, herkese araştırmalarda başarılı olma şansı verme konusunda duyarlıdır	1	2	3	4	5

28.	Okul yönetimi, herkese öğretim alanında başarılı olma şansı verme konusunda duyarlıdır	1	2	3	4	5
29.	Okul yönetimi, herkese yerine getirmesi gereken görevlerinde başarılı olma şansı verme konusunda duyarlıdır	1	2	3	4	5
30.	Okul yönetiminin, öğretmenlerle işbirliği içinde araştırma yapma konusundaki yaklaşımından memnunum	1	2	3	4	5
31.	Yürüttüğüm öğretim faaliyetlerinin, toplumun bir parçası olma şansına erişimimde önemli bir etkisi olduğunu hissediyorum.	1	2	3	4	5
32.	Okulumda yönetimin öğretimsel konularda öğretmenlerle çalışma biçiminden memnunum	1	2	3	4	5
33.	Yaptığım çalışmaların kariyerim için sağladığı fırsatlardan memnuniyet duyuyorum	1	2	3	4	5

Note (s): The scale can be employed in academic studies by following proper citation rules. It is not necessary to obtain permission from the author for its use.

The validity and reliability study of the theory of mind inventory-2 (TOMI-2) Turkish version

Canan Keleş Ertürk^{1*}, Kezban Tepeli²

¹KTO Karatay University, Vocational School of Health Services, Child Development Program, Konya, Türkiye

²Selçuk University, Faculty of Health Sciences, Child Development, Konya, Türkiye

ARTICLE HISTORY

Received: Sep. 26, 2023

Accepted: June 18, 2024

Keywords:

Theory of mind,
Preschool period,
False belief,
Validity,
Reliability.

Abstract: This study aims to conduct the Turkish adaptation, validity, and reliability study of the Theory of Mind Inventory-2 (TOMI-2) developed by Hutchins and Prelock (2016) for 3-5-year-old children. The study group consists of 310 mothers with children in the 3-5 age group in Konya city center. Personal Information Form and Theory of Mind Inventory-2 (TOMI-2) were used as data collection tools in the study. After the TOMI-2 was translated into Turkish, the normality assumption was checked with the "Shapiro-Wilk" test. The relationship between two continuous variables was evaluated with the Pearson Correlation Coefficient. Exploratory Factor Analysis, Confirmatory Factor Analysis, Content Validity, Criterion Validity, and Reliability analyses were also used in the study. The findings of the analyses show that the Turkish version of the TOMI-2 is a valid and reliable measurement tool for children aged 3-5, with 60 items in the original form.

1. INTRODUCTION

Theory of mind (ToM), defined as the ability to predict and explain people's behavior, is considered an important milestone in social cognitive development (Slaughter & Repacholi, 2003). Theory of mind, which also means the capacity to interpret, predict, and explain the behaviors of others according to their underlying mental states, begins to develop from early childhood (Scholl & Leslie, 1999). As theory of mind involves both explaining one's actions and interpreting and predicting the actions of other individuals, it forms the basis for understanding human behavior (Astington & Dack, 2008). Theory of mind refers not only to a cognitive tool used to predict and explain action but also to a system of ideas about mental states and activities (Sodian, 2005, p.112).

Different views and theories on the development of the theory of mind have been developed. According to the theory, the believed situation creates a biased effect on perception and the experience shapes the theory of mind (Flavell, 1999). According to the modular theory, theory of mind is acquired through neurological processes, but performance and experience are not ignored (Sodian & Kristen, 2010). Simulation theory, on the other hand, focuses on knowledge

*CONTACT: Canan KELEŞ ERTÜRK ✉ canankeles90@gmail.com 📍 KTO Karatay University, Vocational School Of Health Services, Child Development Program, Konya, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

about how to perform practical skills (Ratcliffe, 2007). There are two main views of the Theory of Mind. According to the traditional view, ToM is a unifying construct about the cognitive aspects of knowing what another person knows. Assessments here focus on understanding cognition, thinking about what someone thinks, knows, or believes. In the modern view, the theory of mind is a construct that is closely related to language. However, it is not a unifying construct. With the increase in brain imaging studies, there is evidence that ToM has different dimensions such as cognitive ToM, emotional-cognitive ToM, and emotional empathy. In addition, the interpersonal theory of mind, which is explained as thinking about others' thoughts and emotions, and the personal theory of mind, which is explained as thinking about one's thoughts and feelings, involve different neurophysiology and different skill groups (Westby & Robinson, 2014).

A developed theory of mind enables an individual to understand that behavior can be guided by mental states (such as desires, knowledge, and beliefs). Theory of mind is also accepted as a fundamental skill for social cognition (interacting with other individuals) and reading comprehension. In this respect, the development of the theory of mind, which is a comprehensive concept, proceeds in certain stages (Table 1) (Tucci, 2023).

Table 1. *Developmental sequence.*

Stage	Age of Mastery (Months)	Task Description
Diverse Desires Stage	36-48 months	The child is given a choice of two snacks. The child picks a favorite snack. Another character chooses the opposing snack as his/her favorite. The child is asked what the character will choose to eat. The child must inhibit his/her desire and choose the opposing snack.
Diverse Beliefs Stage	36-48 months	The child is given a choice of two locations for a missing cat. The child picks the location where he thinks the cat is hiding. Another character chooses the opposing location. The child is asked where the character will look for the cat. The child must inhibit his/her desire and choose the opposing location.
Social Pretend Stage	48-54 months	The child and assessor pretend to paint a blue cup green. The child is asked what color another character thinks the cup is. The child should say the initial color of the cup.
Knowledge Access	55 months	The child is shown a nondescript box. A toy is hidden inside the box. The child is asked what the character thinks is inside the box. The child must say the character doesn't know.
False Belief-Unexpected Contents	60 months	The child is shown a candy box and when asked what is inside, is expected to answer candy. When the inside of the box is shown, it is understood that it is something different. The child is asked what another character thinks is inside the box and is expected to answer candy.

Babies can distinguish between the movements of animate and inanimate objects around 6 months. Perception-goal psychology, the most basic form of theory of mind, emerges around 9 months. It allows individuals to understand that they may have different perceptual perspectives and different goals and act accordingly. At around 12 months, the development of joint attention begins, and by 18 months, the theory of mind manifests itself in the ability of joint attention. Improvement of basic theory of mind skills occurs between 1-3 years of age. From the age of 4, it emerges in the form of belief-desire psychology. This is associated with the subjective representation of others' ways of seeing the world, which may be inaccurate and incompatible with one's subjective view. Higher-level theory of mind is improved until adulthood (Rakoczy, 2022; Şahin et al. 2019).

In this gradual development process of the theory of mind, evaluation is an important issue. In the assessment of the theory of mind, the false belief task, which involves obtaining accurate

predictions about another person's behavior by referring to that person's mental (false) representation (i.e. their false belief), is used (Slaughter & Repacholi, 2003). The false belief task is implemented in two general forms: unexpected content and unexpected location. The unexpected location task is related to the changing location of the object. The unexpected content is about the contents of a box. Here, the child is shown that there is a very different content (such as money, ribbon) in a box that typically belongs to one type of content (e.g., candy, paint box) and is asked what another person who has not seen the inside of the box thinks is in the box (Miller, 2016, p. 9). In the process of evaluating the Theory of Mind (ToM), it should be taken into consideration that ToM is influenced by social conversation experiences, interactions with siblings, participation in imaginary games, and secure attachment, in short, environmental factors and genetic foundations are also involved in this process (Zufferey, 2010, p.39; Wellman, 2014, p.107; Slaughter et al., 2015; Wellman, 2017; McElwain, et al., 2019).

In the literature, scale development and adaptation studies have been carried out to evaluate the Theory of Mind skills of 3-5-year-old children. Gözün Kahraman (2012) conducted a Turkish adaptation study of the Theory of Mind Scale developed by Wellman and Liu (2004). The scale consists of 6 tasks. Each task is presented with scenarios written for children, small toy figures and pictures are used, and then the child is asked the relevant question. Kılıç Tülü and Ergül (2022) developed the “Theory of Mind Test for 3-5 Year Old Children”. The test includes 27 items for the 3 and 4-year-olds and 26 items for the 5-year-olds. The test is applied by telling short stories about the skills to be measured, showing the photographs, asking the relevant questions, and getting answers from the child. Altıntaş (2014) and Keleş Ertürk & Tepeli (2023) carried out the Turkish adaptation study of the Theory of Mind Task Battery (TOMTB), which was developed and revised by Hutchins and Prelock in 2010. The TOMTB is in booklet format with a test of 15 basic questions, colorful pictures, and accompanying text. The text is read and the child is asked to find the picture showing the correct answer. When the scales developed and adapted in the national literature are examined, it is seen that the assessment of ToM is a developing subject, and scale applications are limited to 3–5-year-old children. This study was planned based on the idea that evaluating the theory of mind across a wider age range and developmental stages, with input from parents, would be more useful. For this reason, the adaptation study of the TOMI-2 will provide a more detailed evaluation of ToM and provide guiding information for national inventory development.

1.1. Present Study

When the developed and adapted measurement tools are examined, it is seen that theory of mind is an emerging topic in the literature and the measurement tools are limited to children. This study is an inventory adaptation study that emerged due to the need for the evaluation of the theory of mind in the field. The TOMI-2, whose Turkish adaptation, validity, and reliability study was conducted within the scope of this study, examines ToM skills in a wider age range and progressively based on parental opinions. This detailed examination is provided by the 6 subscales of the TOMI-2. The Early subscale assesses ToM abilities that typically emerge in late infancy and childhood. The Basic subscale assesses ToM abilities that typically emerge during the preschool years. The Advanced subscale assesses ToM abilities that typically emerge in late childhood but persist into adolescence. The Emotion Recognition subscale focuses on the ability to recognize various emotions. The Comprehension of Mental State Terms subscale provides an understanding of mental state terms. The Pragmatics subscale provides an understanding of the pragmatic and metalinguistic aspects of language (Hutchins & Prelock, 2016).

With the TOMI-2, both the ToM skills of children can be determined, and individually delayed or advanced ToM skills can be revealed and suggestions and points that need to be developed can be determined individually. From this point of view, the TOMI-2 can assess the ToM of 3-5-year-old children more comprehensively and can also be used in clinical assessment.

In light of this information, this study aimed to conduct a scientifically accurate Turkish adaptation, validity, and reliability study of the Theory of Mind Inventory-2 (TOMI-2) developed by Hutchins and Prelock (2016) for 3-5-year-old children. In line with this general purpose, the following sub-goals were tested.

- I. Does the Theory of Mind Inventory-2 (TOMI-2) provide content validity for 3–5-year-old children?
- II. Does the Theory of Mind Inventory-2 (TOMI-2) provide construct validity for 3–5-year-old children?
- III. Does the Theory of Mind Inventory-2 (TOMI-2) provide reliability for 3–5-year-old children?
- IV. Does the Theory of Mind Inventory-2 (TOMI-2) provide criterion validity for 3 to 5-year-old children?

2. METHOD

This study aimed to adapt, validate, and test the reliability of the Theory of Mind Inventory-2 (TOMI-2) for Turkish children aged 3-5, originally developed by Hutchins and Prelock (2016), using the general survey model. There are steps to be followed for a measurement tool developed in one language to be used in another language. According to Hambleton and Patsula (1998), when the purpose of the adapted test is cross-cultural or international assessment, an adapted test is the most effective way to produce an equivalent test in a second language. Considering this principle, the adaptation study of the TOMI-2 was planned.

As stated by Hambleton and Patsula (1999), there are basic principles that should be followed in the process of adapting a measurement tool. To ensure linguistic equivalence, people who are fluent in both languages and have knowledge of the subject should be selected and forward and backward translation techniques should be used. A different group of translators should then review the adapted test. After the linguistic process, a pilot study should be conducted with a small group. After all the arrangements are completed, the application should be made in the sample group and the necessary analyses should be made. In this study, an adaptation study was conducted according to the principles determined by Hambleton and Patsula (1999).

2.1. Participants

The sample of the study was determined by the Appropriate Case Study Group, which is one of the Purposeful Study Groups. A Convenient Case Study Group is the easy selection of individuals and groups to be researched (Sönmez & Alacapınar, 2018: 175). Based on this, the study group of the research consists of a total of 310 children in the 3-5 age group and their mothers attending 5 kindergartens affiliated with the Ministry of National Education in Konya city center. The sample size was estimated based on relative criteria such as the number of items or factors. The sample size for factor analysis was reported as 100=*poor*, 200=*adequate*, 300=*good*, 500=*very good*, and 1000 and above=*excellent*. Bryman and Cramer's sample size recommendation is to apply the number obtained by multiplying the number of items by 5 or 10 (Çokluk et al. 2018). Therefore, the sample in this study was determined as 310 people. Descriptive statistics regarding the personal characteristics of the children are presented in [Table 2](#), and descriptive statistics regarding the characteristics of the parents are presented in [Table 3](#).

[Table 2](#) shows that the average age of the children of the parents who participated in the study was 55.95 months. Of the children, 151 (%48.7) were girls and 159 (%51.3) were boys. 160 (%51.6) were the first child and 72 (%23.2) were the only child. The duration of preschool attendance was less than 6 months for 161 (%51.9) children.

Table 2. Descriptive statistics on the characteristics of children.

	Statistics
Age (Month)	
Mean±SD	55.95±8.28
Min-Max	37-70
Age (Month) category	
37-48 Month	66 (%21.3)
49-60 Month	100 (%32.25)
61-70 Month	144 (%46.45)
Gender	
Female	151 (%48.7)
Male	159 (%51.3)
Birth order	
First child	160 (%51.6)
Middle child or one of the middle children	38 (%12.3)
Last Child	112 (%36.1)
Number of Siblings	
0	72 (%23.2)
1	164 (%52.9)
2	56 (%18.1)
3 and more	18 (%5.8)
Duration of Preschool Education	
0-6 months	161 (%51.9)
7-12 months	47 (%15.2)
13-18 months	35 (%11.3)
19-24 months	36 (%11.6)
More than two years	31 (%10)

Summary statistics are given as *mean ± standard and Median (minimum. maximum)* for numerical data and *Number (Percentage)* for categorical data.

Table 3 shows that while the mothers of 70 (%22.6) children are 29 years old or younger, there are 30 (%9.7) children whose fathers are 29 years old and younger. There are 25 (%8.1) children whose mothers have postgraduate degrees and 50 (%16.1) children whose fathers have postgraduate degrees. In addition, there are 162 (%52.3) children whose mothers are working and 304 (%98.1) whose fathers are working. Of the 88 mothers who selected others (%28.4), 52 were health personnel, 11 were lawyers and 25 were engineers. Of the 132 fathers who selected Other (%42.6), 47 were health personnel, 7 were lawyers, 22 were security personnel, 43 were merchants and 13 were engineers.

Table 3. Descriptive statistics of the characteristics of the parents.

	Statistics
Mother's age	
29 years and below	70 (%22.6)
30-39 years	198 (%63.9)
40-49 years	42 (%13.5)
Father's age	
29 years and below	30 (%9.7)
30-39 years	191 (%61.6)
40-49 years	81 (%26.1)
50 years and older	8 (%2.6)
Mother's Education	
Primary and secondary school	36 (%11.6)
High School	53 (%17.1)
University	196 (%63.2)
Postgraduate	25 (%8.1)
Father's Education	
Primary and secondary school	21 (%6.8)
High School	50 (%16.1)
University	189 (%61)
Postgraduate	50 (%16.1)
Mother's employment status	
Working	162 (%52.3)
Not working	148 (%47.7)
Father's employment status	
Working	304 (%98.1)
Not working	6 (%1.9)
Mother's occupation	
Housewife	132 (%42.6)
Officer	79 (%25.5)
Worker	6 (%1.9)
Self-employed	5 (%1.6)
Other	88 (%28.4)
Father's occupation	
Officer	86 (%27.7)
Worker	22 (%7.1)
Self-employed	70 (%22.6)
Other	132 (%42.6)

Summary statistics are given as *Number (Percentage)* values.

2.2. Data Collection Tools

2.2.1. Personal information form

In the study, the "Personal Information Form" prepared by the researcher was used to determine the demographic characteristics of the parents of children in the 3-5 age group. This form consists of multiple-choice questions about the child's gender, birth order, date of birth, number of siblings, duration of preschool attendance, socio-economic level of the family, parent's age, education level, occupation, and employment status.

2.2.2. Theory of mind inventory-2 (TOMI-2)

The Theory of Mind Inventory is designed to assess social cognitive states. The inventory is completed by parents or individuals primarily responsible for the care of typically developing children between the ages of 2 and 12 and individuals diagnosed with autism spectrum disorder. The first version of the TOMI consists of 42 items. Each item is answered with a 20 cm continuum supporting the statements "definitely no, probably no, undecided, probably, definitely". The participant is asked to read the item and mark the appropriate point on the 20 cm line. The validity and reliability study of the first version of the TOMI was conducted with the participation of 124 participants. The test-retest $r=0.89$; standard error of measurement 1.50; internal consistency Cronbach's Alpha value $\alpha=0.98$; criterion validity $r=0.73$ were calculated for the first version of TOMI. As a result of the analyses, the first version of the TOMI was found to be a valid and reliable measurement tool. Then, the number of items was increased and the 60-item TOMI-2 was created. The norm study of TOMI-2 consists of 802 participants. In the analysis conducted for the structural validity of TOMI-2, the Pearson correlation was found to be $r = 0.67$ ($p < 0.001$). TOMI-2 explains 80% of the total variance. The correlation between TOMI-2 and TOMI is $r=0.89$. TOMI-2 Cronbach Alpha reliability coefficient is $\alpha = 0.98$. For TOMI-2, the standard error of measurement (SEM) was 2.12 for the composite score ($M = 100$, $SD = 15$) and 1.4 for the subscale scores ($M = 50$, $SD = 10$). As a result of the analyses, it was determined that TOMI-2 is a reliable and valid measurement tool. The ToMI-2 consists of 6 subscales and a total of 60 items. Each of the 60 items that make up the ToMI-2 belongs to one of 6 empirically derived subscales (Early, Basic, Advanced, Emotion Recognition, Mental State Term Understanding, and Pragmatic) that reflect a progression in ToM development. The Early subscale focuses on ToM abilities that typically emerge during late infancy and toddlerhood. The Basic subscale includes ToM abilities that typically emerge during the preschool years. The Advanced subscale includes ToM abilities that typically emerge in late childhood but persist into adolescence. The Emotion Recognition subscale includes the ability to recognize various emotions. The Understanding Mental State Terms subscale includes an understanding of mental state terms such as 'want', 'think', and 'know'. The Pragmatics subscale includes understanding the pragmatic and metalinguistic aspects of language. Both manual (paper and pencil) scoring and computer-based scoring can be done with the ToMI-2. In manual (paper-and-pencil) scoring, each of the 60 items that make up the ToMI-2 is scored using the ruler on the last page of the scale. The 20-centimeter ruler gives possible scores ranging from 0 to 20 for each item. Computer-based scoring is accomplished by entering the scores obtained on the TOMI-2 online. By scoring the TOMI-2, raw scores, percentiles, and standard scores can be generated. The examination of raw scores can be useful when the user is interested in individual item-level and/or subscale-level analyses. Percentiles are also obtained in computer-based scoring. Percentiles are a type of ordinal norm-referenced scores. For example, for very young children in early developing ToM capacity, a 2-point difference may result in a relatively large change in percentile rank, whereas for an older child, a 2-point difference may result in a very small change in percentile rank. A standard score is a raw score converted into a scale with known characteristics (e.g., a specific mean and standard deviation). The ToMI-2 uses two different standard scores: the standard score for the composite (overall) score has a mean of 100 and a standard deviation of 15, and the standard scores for the six subscale scores (Early, Basic, Advanced, Emotion Recognition, Mental State Term Understanding, Pragmatics) have a mean of 50 and a standard deviation of 10 (i.e., these are T scores) (Hutchins & Prelock, 2016; Prelock, Hutchins & Bonazinga Bouyea, 2016). Adaptation studies of the original TOMI and TOMI-2 have been conducted with different samples. The adaptation study of the French version of the original TOMI was conducted by Houssa, Mazzone, & Nader-Grosbois (2014) with 107 typically developing children aged 3-5 years. The factor validity study of the TOMI-2 was conducted by Lee et al. (2023) with 420 typically developing children aged 3-7 years in a Taiwanese sample.

2.3. Procedure and Data Analysis

In the translation of TOMI-2 from English to Turkish, forward and backward translation procedure was applied and language equivalence was ensured. Then, field experts were consulted to evaluate the content of the TOMI-2.

In the study, the SPSS software was used to conduct an explanatory factor analysis on the collected data set. In exploratory factor analysis, the dimensions obtained as a linear combination of observed variables are called factors. The factors are hypothetical variables formed by observed variables (Rencher, 2002). To evaluate the suitability of the data for factor analysis, the correlation matrix should be examined. If a significant portion of the coefficients in the correlation matrix is not greater than 0.30, the application of factor analysis may not be appropriate (Hair et al., 1998). The rejection of the basic hypothesis indicates that the variables are suitable for factor analysis.

In addition, the Kaiser-Meyer-Olkin (KMO) criterion, which is obtained by using correlation and partial correlation coefficients, is important in evaluating the suitability of the data for factor analysis. KMO, which is the sample adequacy criterion, takes a value between 0-1. If the KMO value is less than 0.5, the data set is not suitable for factor analysis (Cerney & Kaiser, 1997). In the study, the principal components method was used to obtain the factors. In determining the appropriate number of factors, factor selection criteria as much as the number of eigenvalues greater than one were taken into account. In addition, factor rotation was performed to clarify the variables contributing to the formation of each common factor. The Varimax method was applied to this process. Confirmatory factor analysis was also applied to test the suitability of the factors obtained by exploratory factor analysis to hypothetical or theoretical factor structures. Exploratory factor analysis is generally applied before measurement tool development and construct validity testing.

Confirmatory factor analysis, on the other hand, is used to confirm the structure obtained as a result of explanatory factor analysis or the theoretical factor structure (Brown, 2015). In explanatory factor analysis, the appropriate number of factors to define the basic structure is revealed based on the data matrix, while in confirmatory factor analysis, the number of factors is known a priori. SPSS and Amos package programs were used for confirmatory factor analysis in the study.

Descriptive statistics for the variables in the study were given as number of units (n), percentage (%), mean \pm standard deviation, median (M), minimum (min), and maximum (max) values. In addition, the normality assumption, one of the prerequisites of parametric tests, was examined with the "Shapiro-Wilk" test. The relationship between two continuous variables was evaluated with Pearson Correlation Coefficient. $p < 0.05$ level was considered statistically significant.

2.4. Ethical Principles

This study was conducted by scientific ethical principles. First of all, the developers of the TOMI-2 were contacted and the necessary permissions were obtained. The informed consent form was given to the participants of the study and their participation was ensured voluntarily. It was approved with decision number 2023/043 of KTO Karatay University Non-Pharmaceutical and Medical Device Research Ethics Committee that the study could be carried out.

3. FINDINGS

The mean scores of the items in the TOMI-2 are presented in [Figure 1](#) and the descriptive statistics are in [Table 4](#). When [Figure 1](#) is analyzed, it is seen that 60 items in TOMI-2 have a value between 0 and 20 points. The mean values of the items are shown in the figure. While the 8th item has the highest mean, the 19th item has the lowest mean.

According to [Table 4](#), the mean of the Early subscale in the first part of the TOMI-2 was 14.36 ± 2.44 , the mean of the Basic subscale was 13.38 ± 2.46 and the mean of the Advanced

subscale was 11.00 ± 2.43 points. The mean of the Emotion Recognition Subscale in the second part was 13.35 ± 2.52 , the mean of the Mental State Term Comprehension Subscale was 14.3 ± 3.23 and the mean of the Pragmatic Subscale was 11.34 ± 2.72 points. There are high-level statistically significant relationships between the Early, Basic, Advanced, Emotion Recognition, Mental State Term Comprehension, and Pragmatics subscales in the first and second parts.

Figure 1. Mean score table of the items in TOMI-2.

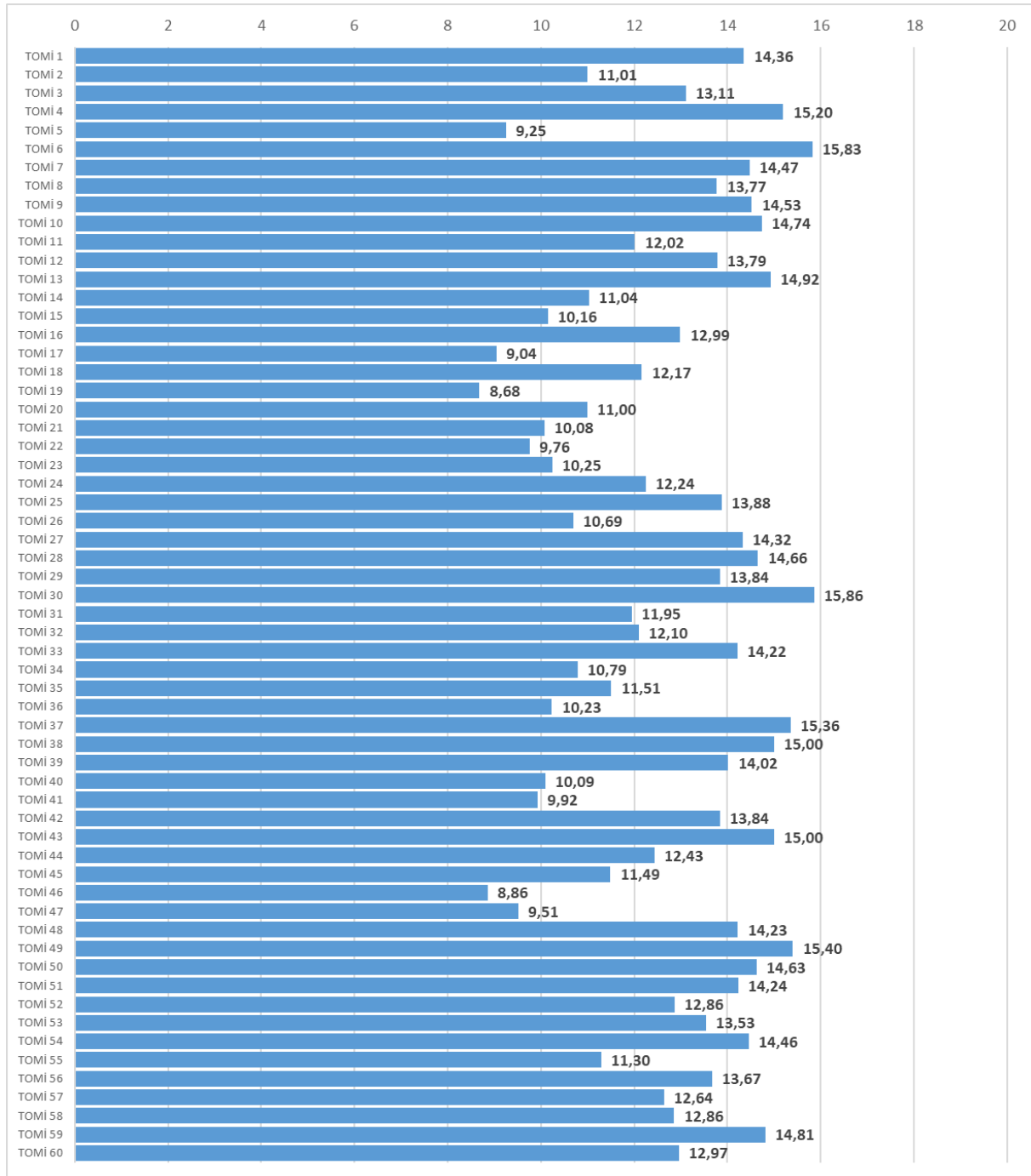


Table 4. Descriptive statistics of TOMI-2.

	Statistics	Early Subscale	Basic Subscale	Advanced Subscale	Emotion recognition Subscale	Mental State Term Comprehension Subscale
Early						
<i>Mean±SD</i>	14.36±2.44					
<i>M (min-max)</i>	14.9 (4-19)					
Basic						
<i>Mean±SD</i>	13.38±2.46	<i>rho</i> =0.871				
<i>M (min-max)</i>	13.6 (4-20)	<i>p</i> <0.001				
Advanced						
<i>Mean±SD</i>	11.00±2.43	<i>rho</i> =0.643	<i>rho</i> =0.744			
<i>M (min-max)</i>	10.9 (3-18)	<i>p</i> <0.001	<i>p</i> <0.001			
Emotion recognition						
<i>Mean±SD</i>	13.35±2.52	<i>rho</i> =0.882	<i>rho</i> =0.826	<i>rho</i> =0.763		
<i>M (min-max)</i>	13.5 (4-18)	<i>p</i> <0.001	<i>p</i> <0.001	<i>p</i> <0.001		
Mental State Term Comprehension						
<i>Mean±SD</i>	14.3±3.23	<i>rho</i> =0.789	<i>rho</i> =0.877	<i>rho</i> =0.566	<i>rho</i> =0.728	
<i>M (min-max)</i>	15.1 (4-36)	<i>p</i> <0.001	<i>p</i> <0.001	<i>p</i> <0.001	<i>p</i> <0.001	
Pragmatics						
<i>Mean±SD</i>	11.34±2.72	<i>rho</i> =0.612	<i>rho</i> =0.73	<i>rho</i> =0.899	<i>rho</i> =0.659	<i>rho</i> =0.558
<i>M (min-max)</i>	11.5 (3-18)	<i>p</i> <0.001	<i>p</i> <0.001	<i>p</i> <0.001	<i>p</i> <0.001	<i>p</i> <0.001

rho: Pearson Correlation Coefficient; Summary statistics are given as *mean ± standard* value. Bolded sections are statistically significant (*p*<0.05).

3.1. Small Group Practice

The Theory of Mind Inventory-2 (TOMI-2) was first administered face-to-face to 20 participants aged 29-49 with children aged 3-5 years. The participants were asked whether the items in the inventory were clearly understood. All participants who participated in the small group application stated that all items in the inventory were clearly expressed and that there was no need for correction.

3.2. Content Validity

For the content validity of the TOMI-2, for which validity and reliability analyses were conducted for parents with 3-5-year-old children, expert opinions were obtained from 5 academicians (2 of them have a bachelor's degree in preschool teaching), 1 with a doctorate in guidance and counseling and 4 with a doctorate in child development and education. All experts reported that the items in the TOMI-2 were necessary and appropriate. Therefore, all items in the original form were used in the data collection process.

3.3. Exploratory Factor Analysis

In this study, exploratory factor analysis was first conducted to assess the construct validity of the TOMI-2. Table 5 shows that the TOMI-2 consists of 2 sections. The first part includes early, basic, and advanced subscales, while the second part includes emotion recognition, mental state term comprehension, and pragmatic subscales. As seen in Table 5, the three-factor structure in the first part explains 64.71% of the total variance, while the three-factor structure in the second part explains 54.91%. The Cronbach Alpha reliability coefficients of the whole inventory and its subscales are also high. The KMO value between 0.90 and 1.00 evaluates the sample adequacy as "very good" (Alpar, 2022, p.625). The Kaiser Meyer Olkin coefficient (KMO) of the TOMI-2 was calculated as 0.94 and the sample was found to be adequate. As a result of the explanatory factor analysis, it is seen that the TOMI-2 is a valid and reliable measurement tool.

Table 5. Exploratory factor analysis results of TOMI-2.

Factor	Item No	Part 1			Explained Variance %	Cronbach Alpha	Factor	Item No	Part 2			Explained Variance %	Cronbach Alpha
		Factor Loads							Factor Loads				
		1	2	3				1	2	3			
Early	3	0.498			14.65	0.930	Emotion Recognition	6	0.643			21.15	0.904
	6	0.683						17	0.481				
	24	0.478						25	0.607				
	25	0.650						48	0.734				
	28	0.624						49	0.775				
	37	0.638						50	0.872				
	38	0.688						32	0.551				
	43	0.746						51	0.741				
	44	0.513						52	0.645				
	48	0.735						55	0.499				
	49	0.803											
	50	0.778											
	Basic	54	0.672						17.22	0.949	Mental State Term Comprehension		
59		0.680			10	0.736							
1			0.641		39	0.691							
4			0.693		33	0.681							
7			0.707		54	0.689							
8			0.590		53	0.636							
9			0.617										
10			0.685										
11			0.710										
12			0.572										
15			0.563										
16			0.550										
26			0.528										
29		0.604											
30		0.678											
31		0.548											
32		0.578											
33		0.612											
Scale											54.91	0.929	
<i>KMO=0.941 Df=300 $\chi^2=4173.421 p<0.001$</i>													

35	0.552		
39	0.648		
42	0.645		
51	0.632		
53	0.574		
57	0.539		
60	0.545		
<hr/>			
2	0.525		
5	0.626		
13	0.512		
14	0.502		
17	0.545		
18	0.479		
19	0.753		
20	0.496		
21	0.614		
22	0.647		
23	0.710		
Advanced	27	0.472	14.84 0.918
	34	0.597	
	36	0.513	
	40	0.533	
	41	0.587	
	45	0.466	
	46	0.620	
	47	0.556	
	52	0.525	
	55	0.524	
	56	0.517	
	58	0.508	
<hr/>			
	Scale	64.71	0.964
<hr/>			
<i>KMO</i> =0.941 <i>Df</i> =1770 $\chi^2=11183.489$ $p<0.001$			

KMO: Kaiser–Meyer–Olkin test; *Df*: Degrees of Freedom

3.4. Confirmatory Factor Analysis

Confirmatory Factor Analysis was conducted with the data obtained from each subscale for the construct validity of the TOMI-2. The boundary values in CFA analysis (Schumacker & Lomax, 2004; Hu & Bentler, 1999; Thompson, 2004; Kline 2015) were evaluated according to Table 6.

Table 6. Boundary values in CFA analysis.

Indices	Boundary Values
χ^2/SD	Perfect $\leq 3 \leq$ Good ≤ 5
RMSEA	Perfect $\leq 0.05 \leq$ Good ≤ 0.08
SRMR	Perfect $\leq 0.05 \leq$ Good ≤ 0.08
CFI	Perfect $\geq 0.95 \geq$ Good ≥ 0.90
NNFI	Perfect $\geq 0.95 \geq$ Good ≥ 0.90
GFI	Perfect $\geq 0.95 \geq$ Good ≥ 0.90
AGFI	Perfect $\geq 0.95 \geq$ Good ≥ 0.90

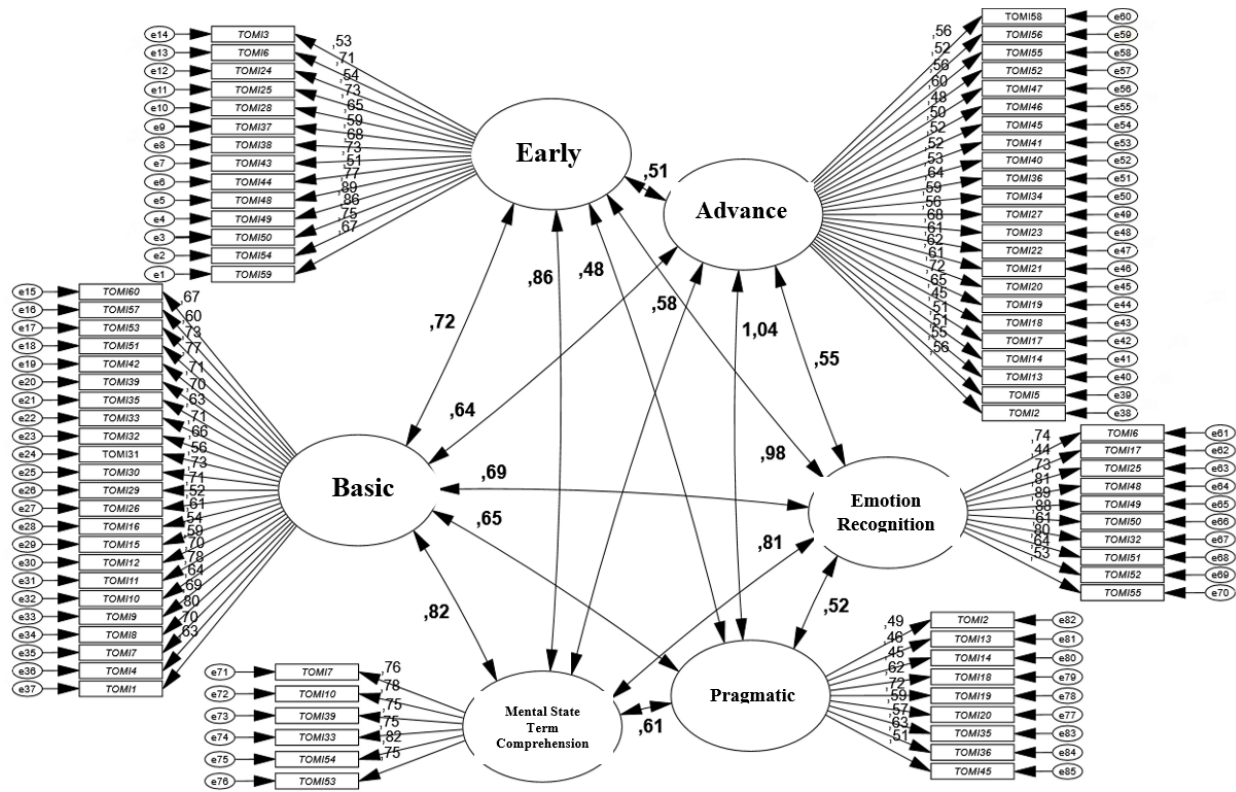
The model ($\chi^2=20043.471$ $df=3450$) obtained as a result of the factor analysis explained in Table 7 includes a total of 6 subscales of the TOMI-2. The fit indices show that the model is an acceptable fit. The first part of the TOMI-2 consists of 60 items and 3 subscales, while the second part consists of 25 items and 3 subscales. The interactions between the two parts and the model created for the TOMI-2 are presented visually in Figure 2.

Table 7. Statistical values for the model fit of TOMI-2.

Measurement	(χ^2/SD)	RMSEA	IFI	CFI	GFI	SRMR
Early	2.184	0.062	0.965	0.964	0.928	0.043
Basic	1.817	0.051	0.947	0.953	0.901	0.047
Advanced	2.209	0.063	0.902	0.901	0.877	0.060
Emotion recognition	3.050	0.080	0.963	0.963	0.938	0.057
Mental State Term Comprehension	2.136	0.061	0.991	0.991	0.981	0.030
Pragmatics	0.964	0.001	0.992	0.999	0.982	0.042

Figure 2 shows the factor loadings of the 6 subscales of the TOMI-2. Accordingly, factor loadings ranged between 0.51 and 0.89 in the Early Subscale, 0.52 and 0.80 in the Basic Subscale, 0.45 and 0.72 in the Advanced Subscale, 0.44 and 0.89 in the Emotion Recognition Subscale, 0.75 and 0.82 in the Mental State Term Comprehension Subscale, and 0.45 and 0.72 in the Pragmatic Subscale.

Figure 2. Confirmatory factor analysis model for the TOMI-2.



3.5. Findings Regarding Criterion Validity

Theory of Mind Task Battery (TOMTB), which is used to evaluate children's theory of mind skills, was used for criterion validity. TOMTB is a 15-item battery that evaluates the theory of mind in 3 subscales: early, basic, and advanced (Keleş Ertürk & Tepeli, 2023). According to Table 8, there is a highly statistically significant relationship between the Early, Basic, Advanced, Emotion Recognition, Mental State Term Comprehension, and Pragmatics subscales of the TOMI-2, the Theory of Mind Task Battery (TOMTB) Early, Basic, Advanced subscales and the TOMTB total score.

Table 8. Findings of the criterion validity of the TOMI-2.

	Early	Basic	Advanced	TOMTB
Early	$\rho=0.327$ $p<0.001$	$\rho=0.227$ $p<0.001$	$\rho=0.193$ $p<0.001$	$\rho=0.295$ $p<0.001$
Basic	$\rho=0.299$ $p<0.001$	$\rho=0.190$ $p<0.001$	$\rho=0.204$ $p<0.001$	$\rho=0.276$ $p<0.001$
Advanced	$\rho=0.188$ $p<0.001$	$\rho=0.158$ $p=0.005$	$\rho=0.161$ $p=0.005$	$\rho=0.215$ $p<0.001$
Emotion Recognition	$\rho=0.291$ $p<0.001$	$\rho=0.218$ $p<0.001$	$\rho=0.142$ $p=0.012$	$\rho=0.254$ $p<0.001$
Mental State Term Comprehension	$\rho=0.252$ $p<0.001$	$\rho=0.183$ $p=0.001$	$\rho=0.252$ $p<0.001$	$\rho=0.293$ $p<0.001$
Pragmatics	$\rho=0.218$ $p<0.001$	$\rho=0.180$ $p=0.001$	$\rho=0.196$ $p<0.001$	$\rho=0.253$ $p<0.001$

3.6. Reliability and Item Analyses of the TOMI-2

The Cronbach Alpha coefficients of the TOMI-2 were calculated as 0.930 for the Early Subscale, 0.904 for the Emotion Recognition Subscale, 0.895 for the Mental State Term Comprehension Subscale, 0.949 for the Basic Subscale, 0.806 for the Pragmatics Subscale and 0.918 for the Advanced Subscale. For the test-retest reliability of the TOMI-2, 30 participants were interviewed again 3 weeks later. According to the results of the analysis, the test-retest reliability (Table 9) ranged between 0.76 and 0.98 ($p < 0.05$). In this case, it can be said that the reliability of the measurements obtained in terms of consistency is good and very good.

Table 9. Test-Retest results of TOMI-2 on a subscale basis.

	Test-retest reliability
Early	0.963
Basic	0.938
Advanced	0.979
Emotion recognition	0.960
Mental State Term Comprehension	0.758
Pragmatics	0.961

CR-Composite Reliability values should be examined for the reliability of the CFA model, and convergent and discriminant validity should be examined for its validity (Çalık et al., 2013). Since the Composite Reliability (CR) value for each factor should exceed 0.7, it can be concluded that the reliability of the CFA model is ensured (Hair et al., 2018). If the CR value is higher than 0.7, it is accepted that the AVE value is greater than 0.4 and it is stated that convergent validity is not impaired (Huang et al, 2013; Fornel & Larcker, 1981; Karadeniz & Kocamaz, 2020; Biçer & Kılıç, 2022). Accordingly, in Table 10, the CR values of the CFA model are between 0.927 and 0.960, while the AVE values are between 0.410 and 0.589. The CR and AVE values prove that the measurement model shows good fit validity.

Table 10. Findings on CR-Composite reliability values.

	N	AVE	CR
Early	14	0.485	0.928
Basic	29	0.458	0.960
Advanced	33	0.410	0.958
Emotion recognition	20	0.519	0.954
Mental State Term Comprehension	12	0.589	0.945
Pragmatics	18	0.419	0.927

The effects between the TOMI-2 items and its subscales are given in in Appendix (see Table A1). Table A1 shows that each of the path coefficients of the subscales in the first part of 60 items is statistically significant ($p < 0.05$). Accordingly, the Early subscale consists of item 3, 6, 24, 25, 28, 37, 38, 43, 44, 48, 49, 50, 54 and 59. The Basic subscale consists of item 1, 4, 7, 8, 9, 10, 10, 11, 12, 15, 16, 26, 29, 30, 31, 32, 33, 35, 39, 42, 51, 53, 57 and 60. The Advanced subscale consists of item 2, 5, 13, 14, 17, 18, 18, 19, 20, 21, 22, 23, 27, 34, 36, 40, 41, 45, 46, 47, 52, 55, 56 and 58. Each of the path coefficients of the subscales in the second part on 25 items is statistically significant ($p < 0.05$). Accordingly, the Early subscale consists of item 3, 6, 24, 25, 28, 37, 38, 43, 44, 48, 49, 50, 54 and 59. The Emotion Recognition subscale consists of items 6, 17, 25, 48, 49, 50, 32, 51, 52 and 55. The Mental State Term Comprehension subscale consists of items 7, 10, 39, 33, 54, and 53. The Pragmatic subscale consists of item 2, 13, 14, 18, 19, 20, 35, 36 and 45. All subscales have a highly statistically significant effect on the item.

The findings regarding the evaluation of the effects between the subscales of the TOMI-2 are given in Table 11. When Table 11 was analyzed, it was found that the relationships between the Early, Basic, Advanced, Emotion Recognition, Mental State Term Comprehension, and Pragmatics subscales of the TOMI-2 were statistically significant.

Item-total correlations for TOMI-2 were also calculated and are given in the appendix (see Table A2). Table A2 shows the item total correlations for the TOMI-2 which range between 0.325 and 0.603. According to Tavşancıl (2002), item-test correlations for the items in the scale are recommended to be 0.30 and above. The values obtained for the TOMI-2 also meet this criterion.

Table 11. Evaluation of the effects between the subscales of the TOMI-2.

			$z\beta$	β	<i>se</i>	<i>t</i>	<i>p</i>
Early	<->	Basic	0.719	5.286	0.712	7.422	<0.001
Basic	<->	Emotion Recognition	0.688	5.439	0.722	7.536	<0.001
Advanced	<->	Emotion Recognition	0.553	3.97	0.634	6.257	<0.001
Basic	<->	Advanced	0.639	5.283	0.814	6.492	<0.001
Advanced	<->	Pragmatics	1.044	7.689	1.079	7.126	<0.001
Early	<->	Advanced	0.508	3.388	0.577	5.867	<0.001
Basic	<->	Mental State Term Comprehension	0.821	7.197	0.883	8.151	<0.001
Advanced	<->	Mental State Term Comprehension	0.583	4.635	0.723	6.413	<0.001
Early	<->	Mental State Term Comprehension	0.858	6.069	0.729	8.325	<0.001
Mental State Term Comprehension	<->	Pragmatics	0.608	4.75	0.712	6.673	<0.001
Basic	<->	Pragmatics	0.646	5.239	0.785	6.671	<0.001
Early	<->	Pragmatics	0.481	3.147	0.55	5.727	<0.001
Emotion Recognition	<->	Pragmatics	0.516	3.637	0.596	6.1	<0.001
Early	<->	Emotion Recognition	0.979	6.252	0.727	8.594	<0.001
Emotion Recognition	<->	Mental State Term Comprehension	0.808	6.149	0.725	8.485	<0.001

β : Regression coefficient, *se*: Standard error, $z\beta$: Standardized regression coefficient. Bolded sections are statistically significant ($p < 0.05$).

4. DISCUSSION and CONCLUSION

Theory of mind forms the basis of the ability to interpret people's communication and actions and also includes the understanding that there are different perspectives (Astington, 2020). The individual also uses the theory of mind when considering the feelings and thoughts of others (Astington & Edward, 2010). Especially in the preschool period, interactions with parents and siblings, and cultural-social-speech experiences shape ToM and can cause significant differences in the developmental stages of ToM (Wellman, 2014; Slaughter et al., 2015). This study was planned to evaluate and support the development process of ToM in preschool and daily life. In this study, the Turkish adaptation, validity, and reliability study of the Theory of Mind Inventory-2 (TOMI-2) developed by Hutchins and Prelock (2016) for 3-5-year-old children was conducted.

It is seen that there is no comprehensive, progressive, and up-to-date assessment of the theory of mind in the national literature, but the national literature also focuses on the development of

the theory of mind. Therefore, it was concluded that adapting the Theory of Mind Inventory-2 to Turkish culture was appropriate. Within the scope of the study, first of all, the necessary permissions for the use of the TOMI-2 were obtained and the process started with its translation into Turkish. Forward and backward translation techniques were used in the translation process, and a different group of translators examined the adapted test. The comprehensibility of the statements was also ensured by conducting a pilot study with a small group. Field experts were also consulted and feedback was received that no item should be removed from the inventory. In the analysis of the data, the normality assumption was examined with the "Shapiro-Wilk" test. The relationship between two continuous variables was evaluated with Pearson Correlation Coefficient. Exploratory Factor Analysis, Confirmatory Factor Analysis, Content Validity, Criterion Validity, and Reliability analyses were also evaluated.

Content validity is determined by applying expert opinion to determine whether the items in the measurement tool are appropriate for measurement (Karasar, 2017, p.195). For this purpose, the opinions of field experts were obtained for the content validity of TOMI-2 and no changes were deemed necessary in the original form.

Kaiser Meyer Olkin coefficient (KMO) was calculated to test the sample adequacy and it was found to be 0.94. A KMO value between 0.90 and 1.00 evaluates the sampling adequacy as "very good" (Alpar, 2022, p.625). When Table 5 is examined, it is seen that TOMI-2 consists of 2 sections. The first part includes early, basic, and advanced subscales, while the second part includes emotion recognition, mental state term comprehension, and pragmatic subscales. The three-factor structure in the first part explains 64.71% of the total variance, while the three-factor structure in the second part explains 54.91%. Generally, an explained variance between 0.50 and 0.70 is considered sufficient. In social sciences, an explained variance between 0.40 and 0.60 is considered acceptable (Alpar, 2022, p.633). In this case, it is possible to evaluate the explained variance of TOMI-2 as sufficient.

The fit indices show that the model has an acceptable level of fit. The factor loadings of the TOMI-2 ranged between 0.51 and 0.89 in the Early Subscale; 0.52 and 0.80 in the Basic Subscale; 0.45 and 0.72 in the Advanced Subscale; 0.44 and 0.89 in the Emotion Recognition Subscale; 0.75 and 0.82 in the Mental State Term Comprehension Subscale; and 0.45 and 0.72 in the Pragmatic Subscale. Factor loadings of 0.60 and above are considered to be high, while loadings between 0.30-0.59 are considered to be moderate (Büyüköztürk, 2002). Accordingly, it can be said that the factor loadings of the TOMI-2 are at high and medium levels.

Criterion validity involves comparing a test that is believed to measure performance, skill, etc., against a standard or another test that measures the same characteristic. (Alpar, 2022, p.536). For this purpose, the Theory of Mind Task Battery (TOMTB) was used for criterion validity, and it was found that there was a highly statistically significant relationship between the TOMI-2 and the Theory of Mind Task Battery (TOMTB).

Reliability, which is also expressed as the stability between independent measurements, can be tested in different ways (Thanasegaran, 2009; Alpar, 2022, p.532). Internal consistency, test-retest reliability, and composite reliability were calculated to test the reliability of the TOMI-2. The Cronbach Alpha reliability coefficients of the TOMI-2 ranged between 0.806 and 0.949, and the test-retest reliability ranged between 0.76 and 0.98. When the Cronbach Alpha reliability coefficient is between 0.60 and 0.79, it is interpreted as highly reliable; when it is between 0.80 and 1.00, it is interpreted as highly reliable (Karagöz, 2019, p.1003). In this case, it can be said that the reliability of the measurements obtained in terms of stability is also highly reliable and highly reliable.

The combined reliability (CR-Composite Reliability) values for the reliability of the CFA model are between 0.927 and 0.960, while the AVE values are between 0.410, and 0.589. The CR

and AVE values prove that the measurement model shows fit validity. Since the CR (Combined Reliability) value should be greater than 0.7 for each factor, it can be said that the reliability of the CFA model is ensured (Hair et al., 2018). If the CR value is greater than 0.7, the AVE value is accepted to be greater than 0.4, and convergent validity is not impaired (Huang cd., 2013; Fornel & Larcker, 1981; Karadeniz & Kocamaz, 2020; Biçer & Kılıç, 2022).

The item-total correlations of the TOMI-2 ranged between 0.325 and 0.603. According to Tavşancıl (2002), item-test correlations for the items in the scale are recommended to be 0.30 and above. The values obtained for TOMI-2 also meet this criterion.

As a result of the analyses, it was determined that the validity and reliability values of the Turkish version of the ToM Inventory-2 with 60 items and 6 subscales by the original model showed an acceptable fit and can be used in Turkish culture. Accordingly, the TOMI-2 can be used to assess children's ToM skills according to parents' views. ToM is a concept that has the power to affect both the social and cognitive development of the individual. The use of the TOMI-2 in children's ToM skills provides a detailed evaluation of ToM. The scores obtained from the inventory reveal at which stage the child is in ToM, and according to the results obtained, guidance can be provided to prepare a supportive environment. However, a limitation of this study is that it was restricted to children aged 3-5 and involved a relatively small sample size. Therefore, it is recommended to conduct validity and reliability analyses of the inventory with different age groups and a larger sample.

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:** KTO Karatay University Non-Pharmaceutical and Medical Device Research Ethics Committee, 2023/043.

Contribution of Authors

Canan Keleş Ertürk: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Kezban Tepeli:** Methodology, Supervision, and Validation.

Orcid

Canan Keleş Ertürk  <https://orcid.org/0000-0001-6247-0073>

Kezban Tepeli  <https://orcid.org/0000-0003-3403-3890>

REFERENCES

- Alpar, R. (2022). *Spor sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik güvenirlik SPSS de çözümlene adımları ile birlikte [Applied statistics and validity and reliability with examples from sports, health and education sciences with analysis steps in SPSS]*. Detay Publishing.
- Altıntaş, M. (2014). Çocuklar için Zihin Kuramı Test Bataryası'nın 4-5 yaş türk çocuklarına uyarlanması, geçerlik güvenirlik çalışması [Adaptation of Theory of Mind Task Battery for Children to 4-5 year old Turkish children, validity and reliability study]. *Master Thesis*, Haliç University.
- Astington, J.W. (2020). The developmental interdependence of theory of mind and language. In N.J. Enfield, & S.C. Levinson (Eds.), *Roots of Human Sociality Culture, Cognition and Interaction* (p.179- 206). Routledge.
- Astington, J.W., & Dack, L.A. (2008). Theory of mind. In M.M. Haith, & J.B. Benson (Eds.), *Encyclopedia of Infant and Early Childhood Development Volume 1* (p. 343-356). Elsevier Inc.
- Astington, J.W., & Edward, M.J. (2010). The development of theory of mind in early childhood. *Social Cognition*.

- Biçer, M., & Kılıç, K.C. (2022). Yönetici Davranışları Ölçeğinin Türkçe'ye uyarlanması: geçerlilik ve güvenilirlik uyarlaması [Adaptation of the Managerial Behaviors Scale into Turkish: validity and reliability adaptation]. *Karamanoglu Mehmetbey University Journal of Social and Economic Research*, 24(42), 277-291.
- Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.
- Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı [Factor analysis: Basic concepts and use in scale improvement]. *Educational Administration in Theory and Practice*, 22, 470-483.
- Cerny, B.A., & Kaiser, H.F. (1977). A study of a measure of sampling adequacy for factor-analytic correlation matrices. *Multivariate Behavioral Research*, 12(1), 43-47.
- Çalık, M., Altunışık, R., & Sütütemiz, N. (2013). Bütünleşik pazarlama iletişimi, marka performansı ve pazarlama performansı ilişkisinin incelenmesi [Analyzing the relationship between integrated marketing communication, brand performance and marketing performance]. *International Journal of Management Economics and Business*, 9(19), 137162.
- Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2018). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve Lisrel uygulamaları* (5. Baskı) [Multivariate statistics for social sciences: SPSS and Lisrel applications (5th Edition)]. Pegem Academy Publishing.
- Flavell, J.H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Reviews Psychology*, 50, 21-45.
- Fornel, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50.
- Gözün Kahraman, Ö. (2012). *Zihin kuramına dayalı eğitim programının 48-60 aylık çocukların bilişsel bakış açısı becerileri ve prososyal davranışları üzerindeki etkisinin incelenmesi* [Investigating the effect of a theory of mind-based education program on cognitive perspective skills and prosocial behaviors of 48-60 month old children] [Doctoral Thesis], Gazi University.
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2018). *Multivariate data analysis* (8. Edition). Cengage Learning.
- Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C., (1998). *Multivariate data analysis*, Prentice Hall.
- Hambleton, R.K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, 45, 153-171.
- Hambleton, R.K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, 1, 1-30.
- Houssa, M., Mazzone, S., & Nader-Grosbois, N. (2014). Validation d'une version francophone de l'inventaire de la Théorie de l'Esprit (ToMI-vf). *European Review of Applied Psychology*, 64, 169-179. <http://dx.doi.org/10.1016/j.erap.2014.02.002>
- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Huang, C.C., Wang, Y.M., Wu, T.W., & Wang, P.A. (2013). An empirical analysis of the antecedents and performance consequences of using the moodle platform. *International Journal of Information and Education Technology*, 3(2), 217-221.
- Hutchins, T.L., & Prelock, P. (2010). *Technical manual for the Theory of Mind Task Battery*. Unpublished Copyrighted Manuscript. Available at: theoryofmindinventory.com
- Hutchins, T.L., & Prelock, P.A. (2016). *Technical manual for the Theory of Mind Inventory-2*. Unpublished Copyrighted Manuscript. Available at: theoryofmindinventory.com.
- Karadeniz, M., & Kocamaz, İ. (2020). An investigation of post-purchase cognitive dissonance and its determinants in online shopping. *Journal of Yasar University*, 15, 307-315.

- Karagöz, Y. (2019). *SPSS-AMOS-META uygulamalı istatistiksel analizler [SPSS-AMOS-META applied statistical analysis]*. Nobel Publishing.
- Karasar, N. (2017). *Bilimsel araştırma yöntemi: Kavramlar ilkeler teknikler* (32. Basım) [*Scientific research method: Concepts principles techniques* (32nd Edition)]. Nobel Publishing.
- Keleş Ertürk, C., & Tepeli, K. (2023). Validity and reliability of Theory of Mind Task Battery (TOMTB) for 3-5 year old children. *International Anatolian Journal of Social Sciences*, 7(3), 627-639. <https://doi.org/10.47525/ulasbid.1322818>
- Kılıç Tülü, B., & Ergül, C. (2022). 3-5 yaş grubu çocuklara yönelik Zihin Kuramı Testi: Geçerlik güvenirlik çalışması [Theory of Mind Test for 3-5 year old children: Validity and reliability study]. *Marmara University Atatürk Education Faculty Journal of Educational Sciences*, 55, 31-61. <https://doi.org/10.15285/maruaebd.966350>
- Kline, R.B. (2015). *Principles and practice of structural equation modeling* (4. Edition). Guilford Publications.
- Lee, S.C., Fu, I.N., Liu, M.R., Yu, T.Y., & Chen, K.L. (2023). Factorial validity of the Theory of Mind Inventory-2 in typically developing children. *Journal of Autism and Developmental Disorders*, 53, 310-318. <https://doi.org/10.1007/s10803-022-05426-0>
- McElwain N.L, Ravindran N., Emery H.T., & Swartz R. (2019). Theory of mind as a mechanism linking mother–toddler relationship quality and child–friend interaction during the preschool years. *Social Development*, 28, 998-1015. <https://doi.org/10.1111/sode.12377>
- Miller, S.A. (2016). *Parenting and theory of mind*. Oxford University Press.
- Prelock, P.A., Hutchins, T.L., & Bonazinga Bouyea, L. (2016). *Theory of Mind Inventory-2 guide to clinical decision-making*. Unpublished Copyrighted Manuscript.
- Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, 1(4), 223-235. <https://doi.org/10.1038/s44159-022-00037-z>
- Ratcliffe, M. (2007). *Rethinking commonsense psychology: A critique of folk psychology, theory of mind and simulation*. Palgrave Macmillan.
- Rencher, A.C. (2002). *Methods of multivariate analysis* (2. Edition), John Wiley & Sons, Inc.
- Scholl, B.J., & Leslie A.M. (1999). Modularity, development, and ‘theory of mind. *Mind & Language*, 14(1), 131-153.
- Schumacker, R. E., & Lomax, R. G. (2004). *A beginner’s guide to structural equation modeling* (2. Edition). Lawrence Erlbaum Associates.
- Slaughter, V., & Repacholi, B. (2003). Introduction: Individual differences in theory of mind what are we investigating?. In B. Repacholi & Virginia Slaughter (Eds.), *Individual Differences in Theory of Mind: Implications for Typical and Atypical Development* (p. 1-13). Psychology Press.
- Slaughter, V., Imuta, K., Peterson, C.C., & Henry, J.D. (2015). Meta-Analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development*, 86(4), 1159-1174. <https://doi.org/10.1111/Cdev.12372>
- Sodian, B. (2005). Theory of mind—the case for conceptual development. In W. Schneider, R. Schumann-Hengsteler & B. Sodian (Eds.), *Young Children’s Cognitive Development Interrelationships Among Executive Functioning, Working Memory, Verbal Ability, And Theory of Mind* (p. 95-130). Lawrence Erlbaum Associates.
- Sodian, B., & Kristen, S. (2010). Theory of mind. In B.M. Glatzeder, V. Goel & A.V. Müller (Eds.), *Towards A Theory of Thinking Building Blocks for A Conceptual Framework* (p. 189-202). Springer.
- Sönmez, V., & Alacapınar, F.G. (2018). *Örneklendirilmiş bilimsel araştırma yöntemleri* (Genişletilmiş 6. Baskı) [*Exemplified scientific research methods* (Expanded 6th Edition)]. Anı Publishing.

- Şahin, B., Bozkurt, A., Usta, M.B., Aydın, M., Çobanoğlu, C., & Karabekiroğlu, K. (2019). Zihin kuramı: Gelişim, nörobiyoloji, ilişkili alanlar ve nörogelişimsel bozukluklar [Theory of mind: Development, neurobiology, related areas and neurodevelopmental disorders]. *Current Approaches in Psychiatry*, 11(1), 24-41. <https://doi.org/10.18863/pgy.390629>
- Tavşancıl, E. (2002). Tutumların ölçülmesi ve spss ile veri analizi [Measurement of attitudes and data analysis with spss]. Nobel Publishing.
- Thanasegaran, G. (2009). Reliability and Validity issues in research. *Integration & Dissemination*, 4, 35-40.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications (1. Edition)*. Psychology Press.
- Tucci, S.L. (2023, July 11). *Theory of mind development and children who are deaf and hard of hearing*. Victorian Deaf Education Institute. [https://www.deafeducation.vic.edu.au/Resource/fact-sheets/Theory of Mind Summary.pdf](https://www.deafeducation.vic.edu.au/Resource/fact-sheets/Theory%20of%20Mind%20Summary.pdf)
- Wellman, H.M. (2017). The development of theory of mind: Historical reflections. *Child Development Perspectives*, 11(3), 207-214. <https://doi.org/10.1111/Cdep.12236>
- Wellman, H.M., & Liu, D. (2004). Scaling of theory of mind tasks. *Child Development*, 75(2), 523 – 541.
- Wellman, H.M. (2014). *Making minds how theory of mind develops*. Oxford University Press.
- Westby, C., & Robinson, L. (2014). A developmental perspective for promoting theory of mind. *Topics In Language Disorders*, 34(4), 362-382. <https://doi.org/10.1097/TLD.000000000000000035>
- Zufferey, S. (2010). *Lexical pragmatics and theory of mind the acquisition of connectives. Volume 201*. Benjamins Publishing Company.

APPENDIX

Table A1. Evaluation of the effects between the items and subscales in the TOMI-2.

			$z\beta$	β	se	t	p
TOMI2_59	←	Early	0.673	1.000			
TOMI2_54	←	Early	0.752	1.178	0.096	12.260	<0.001
TOMI2_50	←	Early	0.857	1.214	0.088	13.748	<0.001
TOMI2_49	←	Early	0.885	1.128	0.080	14.139	<0.001
TOMI2_48	←	Early	0.771	1.138	0.091	12.543	<0.001
TOMI2_44	←	Early	0.514	0.855	0.099	8.633	<0.001
TOMI2_43	←	Early	0.730	0.981	0.082	11.944	<0.001
TOMI2_38	←	Early	0.681	0.911	0.081	11.219	<0.001
TOMI2_37	←	Early	0.590	0.850	0.086	9.829	<0.001
TOMI2_28	←	Early	0.654	0.920	0.085	10.819	<0.001
TOMI2_25	←	Early	0.730	1.052	0.088	11.944	<0.001
TOMI2_24	←	Early	0.541	0.736	0.081	9.070	<0.001
TOMI2_6	←	Early	0.713	0.839	0.072	11.694	<0.001
TOMI2_3	←	Early	0.533	0.695	0.078	8.933	<0.001
TOMI2_51	←	Basic	0.772	1.057	0.086	12.348	<0.001
TOMI2_42	←	Basic	0.707	1.020	0.089	11.431	<0.001
TOMI2_39	←	Basic	0.704	0.997	0.087	11.398	<0.001
TOMI2_35	←	Basic	0.627	0.980	0.095	10.264	<0.001
TOMI2_33	←	Basic	0.709	0.989	0.086	11.459	<0.001
TOMI2_32	←	Basic	0.656	0.984	0.092	10.693	<0.001
TOMI2_31	←	Basic	0.558	0.920	0.100	9.222	<0.001
TOMI2_30	←	Basic	0.730	1.030	0.088	11.764	<0.001
TOMI2_29	←	Basic	0.713	1.029	0.089	11.516	<0.001
TOMI2_26	←	Basic	0.521	0.806	0.093	8.650	<0.001
TOMI2_16	←	Basic	0.613	0.950	0.094	10.059	<0.001
TOMI2_15	←	Basic	0.536	0.785	0.088	8.879	<0.001
TOMI2_12	←	Basic	0.590	0.984	0.101	9.700	<0.001
TOMI2_11	←	Basic	0.698	0.981	0.087	11.303	<0.001
TOMI2_10	←	Basic	0.779	1.140	0.092	12.450	<0.001
TOMI2_9	←	Basic	0.644	0.899	0.086	10.516	<0.001
TOMI2_8	←	Basic	0.694	0.973	0.086	11.255	<0.001
TOMI2_60	←	Basic	0.668	1.000			
TOMI2_57	←	Basic	0.605	0.901	0.079	11.388	<0.001
TOMI2_53	←	Basic	0.732	1.078	0.091	11.795	<0.001
TOMI2_1	←	Basic	0.629	0.848	0.082	10.297	<0.001
TOMI2_4	←	Basic	0.704	0.909	0.080	11.386	<0.001
TOMI2_7	←	Basic	0.795	1.113	0.088	12.674	<0.001
TOMI2_60	←	Basic	0.668	1.000			
TOMI2_57	←	Basic	0.605	0.901	0.079	11.388	<0.001
TOMI2_53	←	Basic	0.732	1.078	0.091	11.795	<0.001
TOMI2_1	←	Basic	0.629	0.848	0.082	10.297	<0.001
TOMI2_4	←	Basic	0.704	0.909	0.080	11.386	<0.001
TOMI2_7	←	Basic	0.795	1.113	0.088	12.674	<0.001

TOMI2_13	←	Advanced	0.509	1.003	0.132	7.595	<0.001
TOMI2_14	←	Advanced	0.508	1.000	0.132	7.585	<0.001
TOMI2_17	←	Advanced	0.449	0.775	0.112	6.892	<0.001
TOMI2_18	←	Advanced	0.648	1.084	0.120	9.019	<0.001
TOMI2_19	←	Advanced	0.716	1.090	0.114	9.601	<0.001
TOMI2_20	←	Advanced	0.608	0.992	0.115	8.636	<0.001
TOMI2_21	←	Advanced	0.617	0.978	0.112	8.727	<0.001
TOMI2_22	←	Advanced	0.609	0.970	0.112	8.643	<0.001
TOMI2_23	←	Advanced	0.682	1.030	0.111	9.317	<0.001
TOMI2_27	←	Advanced	0.555	0.957	0.118	8.106	<0.001
TOMI2_34	←	Advanced	0.590	0.905	0.107	8.456	<0.001
TOMI2_36	←	Advanced	0.636	1.095	0.123	8.907	<0.001
TOMI2_40	←	Advanced	0.533	0.776	0.099	7.869	<0.001
TOMI2_41	←	Advanced	0.523	0.839	0.108	7.758	<0.001
TOMI2_45	←	Advanced	0.521	1.052	0.136	7.731	<0.001
TOMI2_46	←	Advanced	0.495	0.767	0.103	7.440	<0.001
TOMI2_47	←	Advanced	0.481	0.756	0.104	7.281	<0.001
TOMI2_52	←	Advanced	0.597	0.943	0.110	8.535	<0.001
TOMI2_55	←	Advanced	0.565	0.880	0.107	8.202	<0.001
TOMI2_58	←	Advanced	0.562	0.930	0.114	8.171	<0.001
TOMI2_56	←	Advanced	0.525	0.888	0.114	7.775	<0.001
TOMI2_2	←	Advanced	0.556	1.000			
TOMI2_5	←	Advanced	0.546	0.845	0.087	9.754	<0.001
TOMI2_41	←	Advanced	0.523	0.839	0.108	7.758	<0.001
TOMI2_45	←	Advanced	0.521	1.052	0.136	7.731	<0.001
TOMI2_46	←	Advanced	0.495	0.767	0.103	7.440	<0.001
TOMI2_47	←	Advanced	0.481	0.756	0.104	7.281	<0.001
TOMI2_52	←	Advanced	0.597	0.943	0.110	8.535	<0.001
TOMI2_55	←	Advanced	0.565	0.880	0.107	8.202	<0.001
TOMI2_58	←	Advanced	0.562	0.930	0.114	8.171	<0.001
TOMI2_56	←	Advanced	0.525	0.888	0.114	7.775	<0.001
TOMI2_2	←	Advanced	0.556	1.000			
TOMI2_5	←	Advanced	0.546	0.845	0.087	9.754	<0.001
TOMI2_6	←	Emotion recognition	0.737	1.000			
TOMI2_17	←	Emotion recognition	0.440	0.861	0.112	7.692	<0.001
TOMI2_25	←	Emotion recognition	0.726	1.038	0.080	13.049	<0.001
TOMI2_48	←	Emotion recognition	0.813	1.120	0.076	14.765	<0.001
TOMI2_49	←	Emotion recognition	0.892	1.061	0.065	16.375	<0.001
TOMI2_50	←	Emotion recognition	0.879	1.165	0.072	16.114	<0.001
TOMI2_32	←	Emotion recognition	0.612	1.005	0.093	10.858	<0.001
TOMI2_51	←	Emotion recognition	0.796	1.150	0.080	14.421	<0.001
TOMI2_52	←	Emotion recognition	0.638	1.018	0.090	11.346	<0.001
TOMI2_55	←	Emotion recognition	0.534	0.879	0.093	9.398	<0.001
TOMI2_6	←	Emotion recognition	0.737	1.000			
TOMI2_17	←	Emotion recognition	0.440	0.861	0.112	7.692	<0.001
TOMI2_25	←	Emotion recognition	0.726	1.038	0.080	13.049	<0.001
TOMI2_48	←	Emotion recognition	0.813	1.120	0.076	14.765	<0.001

TOMI2_49	←	Emotion recognition	0.892	1.061	0.065	16.375	<0.001
TOMI2_50	←	Emotion recognition	0.879	1.165	0.072	16.114	<0.001
TOMI2_32	←	Emotion recognition	0.612	1.005	0.093	10.858	<0.001
TOMI2_51	←	Emotion recognition	0.796	1.150	0.080	14.421	<0.001
TOMI2_52	←	Emotion recognition	0.638	1.018	0.090	11.346	<0.001
TOMI2_55	←	Emotion recognition	0.534	0.879	0.093	9.398	<0.001
TOMI2_7	←	Mental State Term Comprehension	0.764	1.000			
TOMI2_10	←	Mental State Term Comprehension	0.777	0.995	0.069	14.328	<0.001
TOMI2_39	←	Mental State Term Comprehension	0.748	1.011	0.074	13.689	<0.001
TOMI2_33	←	Mental State Term Comprehension	0.746	0.986	0.072	13.657	<0.001
TOMI2_54	←	Mental State Term Comprehension	0.818	1.108	0.073	15.213	<0.001
TOMI2_53	←	Mental State Term Comprehension	0.748	1.081	0.079	13.695	<0.001
TOMI2_7	←	Mental State Term Comprehension	0.764	1.000			
TOMI2_10	←	Mental State Term Comprehension	0.777	0.995	0.069	14.328	<0.001
TOMI2_39	←	Mental State Term Comprehension	0.748	1.011	0.074	13.689	<0.001
TOMI2_33	←	Mental State Term Comprehension	0.746	0.986	0.072	13.657	<0.001
TOMI2_54	←	Mental State Term Comprehension	0.818	1.108	0.073	15.213	<0.001
TOMI2_53	←	Mental State Term Comprehension	0.748	1.081	0.079	13.695	<0.001
TOMI2_20	←	Pragmatics	0.594	1.000			
TOMI2_19	←	Pragmatics	0.719	1.111	0.105	10.567	<0.001
TOMI2_18	←	Pragmatics	0.624	1.046	0.110	9.533	<0.001
TOMI2_14	←	Pragmatics	0.450	0.882	0.121	7.312	<0.001
TOMI2_13	←	Pragmatics	0.456	0.809	0.109	7.406	<0.001
TOMI2_2	←	Pragmatics	0.494	0.941	0.119	7.921	<0.001
TOMI2_35	←	Pragmatics	0.567	0.981	0.111	8.850	<0.001
TOMI2_36	←	Pragmatics	0.635	1.118	0.116	9.659	<0.001
TOMI2_45	←	Pragmatics	0.513	1.010	0.124	8.172	<0.001
TOMI2_20	←	Pragmatics	0.594	1.000			
TOMI2_19	←	Pragmatics	0.719	1.111	0.105	10.567	<0.001
TOMI2_18	←	Pragmatics	0.624	1.046	0.110	9.533	<0.001
TOMI2_14	←	Pragmatics	0.450	0.882	0.121	7.312	<0.001
TOMI2_13	←	Pragmatics	0.456	0.809	0.109	7.406	<0.001
TOMI2_2	←	Pragmatics	0.494	0.941	0.119	7.921	<0.001
TOMI2_35	←	Pragmatics	0.567	0.981	0.111	8.850	<0.001
TOMI2_36	←	Pragmatics	0.635	1.118	0.116	9.659	<0.001
TOMI2_45	←	Pragmatics	0.513	1.010	0.124	8.172	<0.001

β: Regression coefficient, se: Standard error, zβ: Standardized regression coefficient. Bolded sections are statistically significant ($p < 0.05$).

Table A2. Item total correlations for the TOMI-2.

Factor	Item No	Item Total Correlation
Early	3	0.468
	6	0.540
	24	0.559
	25	0.682
	28	0.633
	37	0.534
	38	0.609
	43	0.650
	44	0.482
	48	0.567
	49	0.634
	50	0.583
	54	0.630
	59	0.619
	Basic	1
4		0.549
7		0.630
8		0.587
9		0.502
10		0.652
11		0.538
12		0.416
15		0.413
16		0.520
26		0.254
29		0.589
30		0.603
31		0.450
32		0.591
33		0.584
35		0.541
39		0.350
42	0.582	
51	0.639	
53	0.605	
57	0.528	
60	0.593	
Advanced	2	0.430
	5	0.435
	13	0.589
	14	0.450
	17	0.337
	18	0.567
	19	0.462
	20	0.467

	21	0.533
	22	0.509
	23	0.587
	27	0.551
	34	0.542
	36	0.570
Advanced	40	0.496
	41	0.488
	45	0.448
	46	0.383
	47	0.442
	52	0.630
	55	0.606
	56	0.583
	58	0.608
	6	0.575
	17	0.388
	25	0.690
	48	0.615
Emotion Recognition	49	0.686
	50	0.657
	32	0.622
	51	0.692
	52	0.627
	55	0.567
	7	0.639
	10	0.639
Mental State Term Comprehension	39	0.371
	33	0.557
	54	0.662
	53	0.643
	2	0.347
	13	0.555
	14	0.325
Pragmatics	18	0.530
	19	0.375
	20	0.436
	35	0.515
	36	0.478
	45	0.378

The effect of rater training on rating behaviors in peer assessment among secondary school students

Nazira Tursynbayeva¹, Umur Öç², İsmail Karakaya^{3*}

¹Khoja Akhmet Yassawi International Kazakh-Turkish University, Faculty of Social and Human Sciences, Department of Pedagogy and Psychology, Turkestan, Kazakhstan

²Ministry of National Education, Refahiye District National Education Directorate, Erzincan, Türkiye

³Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY

Received: Feb. 17, 2024

Accepted: July. 21, 2024

Keywords:

Peer assessment,
Rating behavior,
Rater training,
Writing skills.

Abstract: This study aimed to measure the effect of rater training given to improve the peer assessment skills of secondary school students on rater behaviors using the many-facet Rasch Measurement model. The research employed a single-group pretest-posttest design. Since all raters scored all students, the analyses were carried out in a fully crossed (s x r x c) pattern. There were three facets in the research: student, rater, and criteria. The study group consisted of 25 seventh-grade students at a public school in Ankara in the 2021-2022 academic year. All 25 students in the study group were instructed to write compositions. The compositions were examined by the researchers, and 10 were selected for peer assessment. Before the experiment, students were asked to evaluate their peers' writing skills according to the rubric developed by the researchers. Then, rater training was given to the students for four weeks. After the rater training, the students were instructed to re-evaluate the writing skills of their peers. In the research, four rater behaviors were examined: rater severity, rater leniency, differentiated rater severity, and differentiated rater leniency. When the research results were examined, it was observed that rater training contributed to reducing severity, leniency, and differentiated severity and leniency behaviors.

1. INTRODUCTION

One of the aims of today's education system is to prepare and support students for daily life. Helping students acquire and develop daily life skills is a major objective of curriculum. One of the practices applied as part of these objectives is the observation and assessment of students across the curriculum. Assessment and evaluation are used to measure the learning outcomes, behavior acquisition, and the effectiveness of teaching programs (Ertürk, 1979).

The proper functioning of the evaluation mechanism allows for quick and effective solutions to potential problems in the system. Monitoring student progress becomes easier, and learning outcomes are more easily and accurately identified. In this way, both the quality of education increases and development is ensured in a way to facilitate and promote adaptation to

*CONTACT: İsmail KARAKAYA ✉ ikarakaya@gazi.edu.tr 📧 Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Ankara, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

innovations (Çeçen, 2011; Gürten et al., 2019; İşman & ESKİCUMALI, 2003; Kurudayioğlu et al., 2008; Turgut & Baykul, 2010; Yaşar, 2017).

Teachers use different evaluation methods when examining the effects of the educational process. If the evaluation methods used are independent of the student, the evaluation process will be incomplete for the student. This is because students usually have more information about their peers' tasks than their teachers (Somervell, 1993). Involving students in the assessment process increases teacher-student and student-student interaction and contributes to the development of students' responsibility-taking behaviors (Keaten & Richardson, 1993). One of the assessment approaches involving active student participation in the assessment process is peer assessment. Peer assessment is the evaluation of classmates according to specified criteria (Boud et al., 1999). Peer assessment allows students to work together effectively (Kutlu et al., 2010).

The biggest problem in educational settings where peer assessment is used is the reliability of the scores obtained (Donnon et al., 2013). When appropriate environments and conditions are not provided, students cannot make objective evaluations and this causes the evaluation to produce incorrect results (Ellington et al., 1997). In addition, the validity of the assessment will be negatively affected as there will be a rater effect on the assessment. Some of the common rater behaviors are rater severity, rater leniency, and bias (Myford & Wolfe, 2003). The tendency of a rater to give lower scores than other raters in the rater group is called rater severity, and the tendency to give higher scores is called rater leniency (Myford & Wolfe, 2004). Rater bias is the tendency of the rater to be sometimes harsh and sometimes generous when scoring students, depending on the characteristics of the students other than the measured characteristic (Knoch et al., 2007). To reduce or eliminate these rater behaviors, it is recommended to use rubrics, use more than one rater, and provide rater training (Andrade, 2005; Hauenstein & McCusker, 2017; Kubiszyn & Borich, 2024; Lumley & McNamara, 1995; Oosterhof, 1999). In the current study, all of the suggested methods were used to make the rating more valid and reliable. Peer assessment involves, by nature, more than one rater. In peer assessment, before creating the relevant assessment tool, the basic behaviors and criteria related to the task are identified with the students, and the expected behaviors of the students are listed. Students should be involved from the first stage of the assessment process. The type of assessment to be used and which learning outcomes will be assessed should be well explained to the students beforehand. The tasks should be appropriate to the level of the students, similar approaches should be used frequently in class, assessment criteria should be prepared together with the students, and possible disagreements should be resolved (Alıcı, 2010; Bushell, 2006; Kutlu et al., 2010; Stiggins & Chappuis, 2005; Woolfolk et al., 2008). After the definitions and explanations about the task are completed, the students should be instructed on how the assessment should be done.

After the assessment tool is created, rater training should be provided to support students in rating objectively. Lack of objectivity in rating is one of the biggest problems encountered during implementation (Donnon et al., 2013). Students' involvement in the rating process supports teaching, influences students' rating behaviors, and contributes to the validity and reliability of the rating. Students' tendency to give a higher score to their close friends or to classmates who are at the top of the class, their failure to fulfill the responsibilities that need to be observed during peer assessment, and their inability to fully comprehend the criteria may negatively affect the peer assessment process (May, 2008). Students' subjective rating behavior may lead to a biased evaluation of the learning process and learning outcomes, and students failing to fulfill their tasks fully may come to the forefront. Studies show that in peer assessment, students may resort to different ways to give each other higher scores and that they may be biased (Greenan et al., 1997; Johnson & Smith, 1997). In addition, all kinds of rater effects can be expected in peer assessment (Farh et al., 1991; Heslin, 2005). Examining the effectiveness of the techniques used to increase objectivity in evaluations using peer assessment

is very important for the reliability of the scores obtained and the validity of the inferences to be made based on the scores. Therefore, providing rater training may contribute to rating validity. When the literature is examined, it is seen that several studies found that rater training contributed significantly to rating accuracy (Bijani, 2018; Congdon & MeQueen, 2000; Fahim & Bijani, 2011; Kondo, 2010; Loignon et al., 2017; Martin & Locke, 2022; May, 2008; Yeşilçınar & Şata, 2021).

Eliminating or reducing undesirable rater behaviors in performance assessment will contribute to the validity, accuracy, and reliability of the results. When the literature is examined, it is seen that there are studies that investigate the effect of rater training on rater behavior in peer assessment among groups at university level and above (Loignon et al., 2017; Martin & Locke, 2022; May, 2008; Yeşilçınar & Şata, 2021). However, there is no study that investigates the effect of rater training on rater behaviors in peer assessment among students at secondary school level. To fill this gap, this study was conducted to determine how rater training given to improve the peer evaluation skills of secondary school students affects their peer rating behaviors.

This research is important to determine the rater behaviors that occur during the process of using peer evaluation and to determine the effect of rater training on eliminating or reducing these behaviors. Focusing especially on the rater behaviors of secondary school students in the peer evaluation process shows the originality of the study. In light of all this information, it was aimed to investigate the effect of rater training with the multi-facet Rasch model in order to provide more objective and accurate scoring in the evaluation of the writing tasks prepared by secondary school students. For this purpose, answers were sought to the following questions.

- 1) Regarding peer evaluation scores before rater training;
 - a) What is the severity and leniency of the raters?
 - b) What are the raters' differentiated leniency and severity behaviors?
- 2) Regarding peer evaluation scores after rater training;
 - a) What is the severity and leniency of the raters?
 - b) What are the raters' differentiated leniency and severity behaviors?

2. METHOD

2.1. Study's Design

This study employed a single-group pretest-posttest design, aiming to measure secondary school students' rater behaviors when evaluating the writing skills of peers and the effect of rater training on the students' rating behavior in peer assessment, using the many-facet Rasch measurement model. Since each rater scored all students, the analyzes were carried out in a fully crossed pattern. There were three facets in the research: rater, criterion, and student.

2.2. Study Group

The study group consisted of 25 seventh-grade students at a public school in Ankara in the 2021-2022 academic year. The students included in the study were selected according to the following criteria: not having received rater training before, willingness to participate in the study voluntarily, and traceability.

2.3. Data Collection Tools

A writing task and an analytical rubric developed by the researchers were used as data collection tools in the study. During the analytical rubric development process, opinions were taken from three Turkish teachers and two measurement and evaluation experts. The content validity index of the measurement tool was determined using the Lawshe (1975) technique based on expert opinions (CVR=0.99, $p<0.05$). The criteria were arranged to suit the students' levels, and the analytical rubric was finalized. In the rubric, each criterion was evaluated on a four-point scale (1: very unsuccessful; 4: very successful). After the rubric was finalized, validity and reliability studies were conducted. Exploratory factor analysis was conducted to provide evidence for the validity of the rubric. While conducting exploratory factor analysis, the average of the scores

given by the 25 raters to the students' writing tasks was used. Before proceeding with exploratory factor analysis, assumptions such as sample size, multiple normality, linearity and outliers were examined. Çokluk et al. (2021, p. 206) state that when determining the sample size in exploratory factor analysis, the individual/item ratio should be at least 2:1. In the current study, it was determined that the sample size assumption was met because the student-criterion ratio was greater than 2:1 (25:7). Additionally, Guadagnoli and Velicer (1988) criticized the theoretical relationship between sample size and number of items and conducted a Monte Carlo study. They state that even if the number of samples in their study is less than 50, values with a factor loading of 0.80 or more will be sufficient for the sampling assumption. Considering that the factor loadings in the current study are greater than 0.80 (C1= 0.948, C2= 0.945, C3= 0.949, C4= 0.954, C5= 0.922, C6= 0.942, C7= 0.957). It was determined that the number assumption was met. For the multivariate normality assumption, the univariate normality assumption must first be examined (Çokluk et al., 2021, p. 29). After determining that all variables meet the univariate normality assumption (Shapiro-Wilk: $p_1= 0.77$, $p_2= 0.42$, $p_3= 0.23$, $p_4= 0.10$, $p_5= 0.34$, $p_6= 0.66$, $p_7= 0.10$, multiple normality assumption ($p_{1,2,3,4,5,6,7}>0.05$) was examined. The multiple normality assumption was examined with the help of Scatter Plot Matrix, and it was determined that the multiple normality assumption was met. Providing the multiple normality assumption shows that the relationship between the variables is linear (Büyüköztürk, 2002). Additionally, it was determined that there were no extreme values in the data. After determining that the exploratory factor analysis assumptions were met, the KMO test and Bartlett Sphericity test were performed to determine whether the data were suitable for analysis. The KMO value of the data set was 0.909, and the Bartlett test of sphericity was significant ($p<0.00$). A KMO test value of 0.90 or above is considered excellent (Hutcheson & Sofroniou, 1999). The fact that the Bartlett Test of Sphericity result is statistically significant is another indication that the data set is suitable for exploratory factor analysis (Field, 2005). This shows that the data set is suitable for exploratory factor analysis. Exploratory factor analysis was conducted by taking the average of the scores given by the raters. As a result of exploratory factor analysis, it was found that the criteria were gathered under one factor, and the explained variance was 89.344%. Factor loadings of each criterion were 0.948, 0.945, 0.949, 0.954, 0.922, 0.942, and 0.957, respectively. Additionally, the Cronbach alpha reliability of the measurements was calculated and found to be 0.98. According to all these results, it can be said that the analytical rubric developed in this study provides valid and reliable results.

2.4. Data Collection Process

The study involved a two-stage data collection process. In the first stage, the analytical rubric to be used in writing skill evaluation was prepared and the rater group was informed about peer assessment, the writing task, and the rubric. In addition, sample applications were shared with the rater group. Ten compositions, selected from those written by the students, were distributed to the students for scoring. Students were given 10 minutes for each composition, 100 minutes in total, for scoring. The students evaluated the compositions written by their peers, and pre-test scores were obtained. In the second stage, the students received rater training two hours a week for a total of four weeks, totaling eight class hours, and then the students were asked to score their peers' compositions once again, and post-test scores were obtained.

2.5. Data Analysis

In the study, ten compositions written by 25 students for a task were selected. These tasks were scored by 25 students according to seven criteria. The average of the scores given by 25 students to each criterion was used in the factor analysis. For the multi-facet Rasch model, the scores given by 25 students to 10 writing tasks were used.

Analyses were performed using the FACET package program. Before proceeding with the analysis, the assumptions of the many-facet Rasch model, including unidimensionality, local

independence, and model-data fit, were examined (Eckes, 2011, p. 124; Farrokhi et al., 2012). As a result of exploratory factor analysis, it was seen that the measurement tool was unidimensional. Meeting the unidimensionality assumption also indicates that the local independence assumption is met (Hambleton et al., 1991). For model-data fit, the ratios of the standardized residuals in the ± 2 and ± 3 intervals were examined. Linacre (2014) stated that the proportion of standardized residuals outside the ± 2 interval should not exceed 5%, and the proportion of standardized residuals outside the ± 3 interval should not exceed 1%. In the study, the total number of interactions was 1750 (10 student * 7 criteria * 25 raters), the proportion of standardized residuals outside the ± 2 interval was 4.29% (n=75), and the proportion of standardized residuals outside the ± 3 interval was 0.74% (n=13). As such, it can be said that model-data fit is achieved, and the inferences to be made in line with the analysis results are valid.

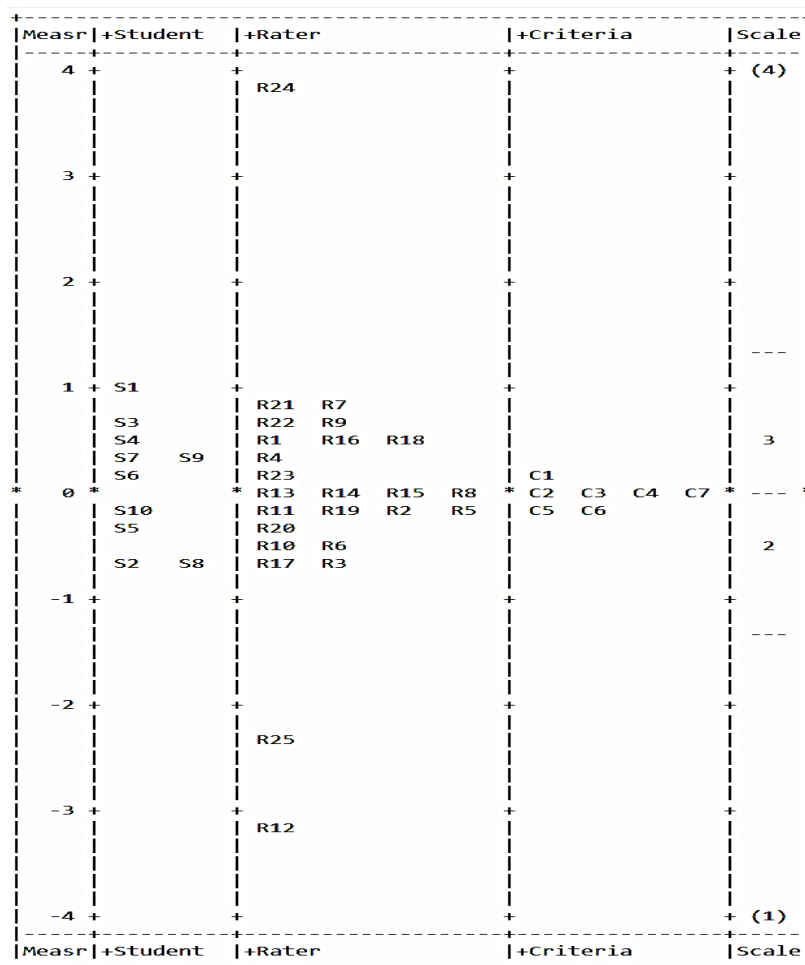
3. RESULTS

The findings obtained in the study are presented under two separate subheadings. The results before rater training (pre-test) are reported under the first, and the results after rater training (post-test) are reported under the second subheading. Under both subheadings, group statistics are given first, followed by individual statistics on a student basis.

3.1. Research Findings Before Rater Training (Pre-Test)

The pre-test calibration map of peer scores, the rater facet measurement report, rater severity and leniency, and biased interactions measured before rater training within the scope of the study are given below.

Figure 1. Calibration map of peer scores before rater training.



When the calibration map of peer scores before rater training in [Figure 1](#) is examined, it is seen that the facets are on a logit scale. A high or low logit value has different implications depending on the relevant facet. In the student column, a high logit value at the top of the column indicates a high level of ability, whereas a low logit value at the bottom indicates a low level of ability. In the rater column, the raters with the highest logit values at the top of the column score leniently, while those with the lowest logit value at the bottom score severely. In the criterion column, a high logit value at the top of the column indicates a highly difficult criterion, whereas a low logit value at the bottom indicates low difficulty. To exemplify, when the calibration map is examined, it is seen that the student with the highest ability level in the pre-test is S1, and the students with the lowest ability levels are S2 and S8. The most lenient rater is R24, while the most severe rater is R12. It is also seen that C5 and C6 are the most difficult criteria, while C1 is the easiest. The fact that the student, rater, and criteria facets take values along the negative and positive ends of the logit scale indicates that the students' ability levels, the criteria difficulty levels, and rater rating behaviors are differentiated. The rater facet measurement reports for a detailed examination of rater behaviors are presented in [Table 1](#).

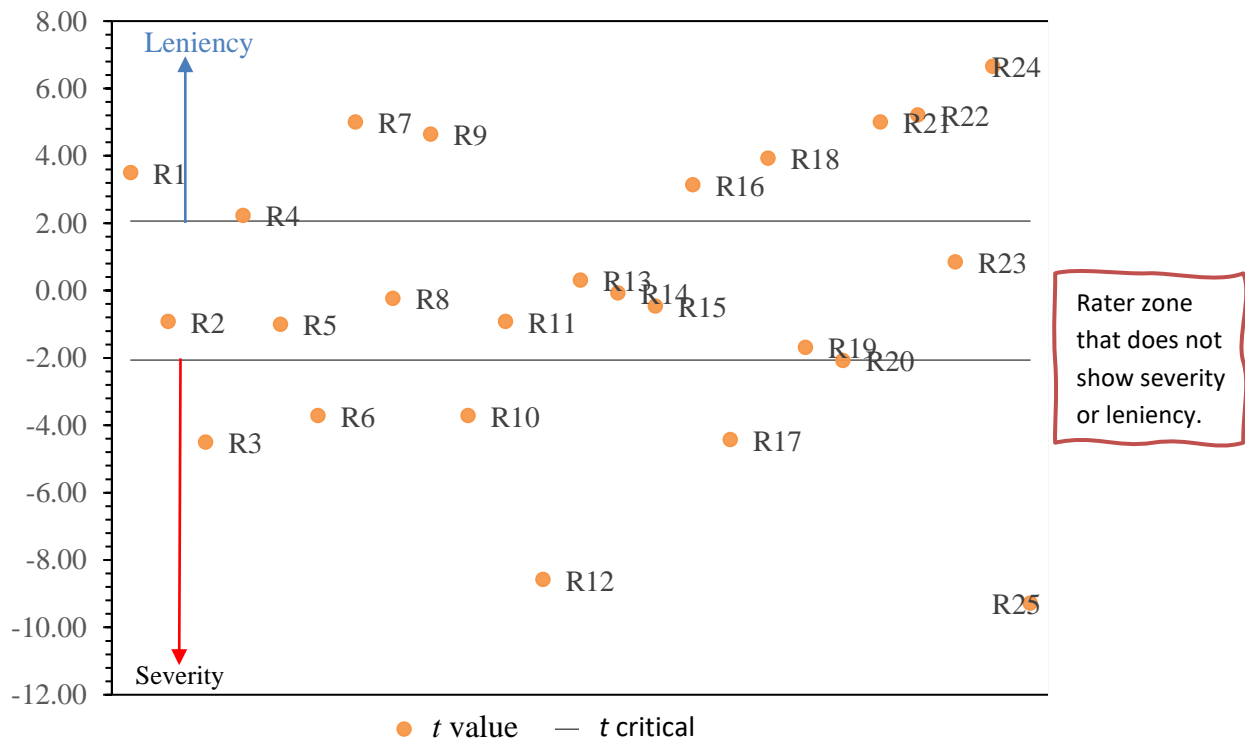
Table 1. Rater facet pre-test measurements measurement report.

Rater	Obsvd Average	Fair Average	Logit	Model S.E.	Infit	Outfit
R1	2.96	3.00	0.49	0.14	0.99	1.00
R2	2.47	2.46	-0.12	0.13	0.82	0.79
R3	2.06	2.01	-0.63	0.14	0.98	1.03
R4	2.80	2.83	0.29	0.13	1.39	1.39
R5	2.46	2.45	-0.13	0.13	1.31	1.27
R6	2.14	2.10	-0.52	0.14	0.80	0.76
R7	3.14	3.20	0.75	0.15	0.79	0.80
R8	2.54	2.54	-0.03	0.13	1.30	1.28
R9	3.07	3.13	0.65	0.14	0.95	1.01
R10	2.14	2.10	-0.52	0.14	0.62	0.69
R11	2.47	2.46	-0.12	0.13	1.23	1.21
R12	1.11	1.10	-3.09	0.36	0.94	1.26
R13	2.60	2.61	0.04	0.13	1.13	1.15
R14	2.56	2.56	-0.01	0.13	0.91	0.89
R15	2.51	2.51	-0.06	0.13	0.79	0.77
R16	2.91	2.96	0.44	0.14	0.63	0.66
R17	2.07	2.02	-0.62	0.14	1.10	1.11
R18	3.00	3.05	0.55	0.14	1.11	1.09
R19	2.39	2.37	-0.22	0.13	1.25	1.26
R20	2.34	2.32	-0.27	0.13	0.99	1.02
R21	3.14	3.20	0.75	0.15	0.82	0.87
R22	3.13	3.19	0.73	0.14	0.85	0.94
R23	2.66	2.67	0.11	0.13	0.83	0.84
R24	3.96	3.96	3.86	0.58	0.94	0.71
R25	1.24	1.21	-2.32	0.25	1.01	1.62
Mean	2.56	2.56	0.00	0.17	0.98	1.02
S (population)	0.58	0.61	1.17	0.10	0.20	0.24
S (sample)	0.60	0.62	1.20	0.10	0.21	0.25
Model, Population			RMSE= 0.19	Adj S.D.= 1.16	Separation= 6.00	
			Strata= 8.33	Reliability= 0.97		
Model, Sample			RMSE= 0.19	Adj S.D.= 1.18	Separation= 6.12	
			Strata= 8.50	Reliability= 0.96		
Model, Chi-square (fixed effect):	433.1	df= 24	p= 0.00			
Model, Chi-square (Normal):	21.3	df= 23	p=0.56			

Table 1 shows the observed and adjusted means, logit values, standard error of logit values, concordance and non-concordance values of the raters before rater training. The logit measures of the raters ranged between -3.09 and 3.86, with a difference of 6.95. A positive value in the logit values of the raters indicates leniency, and a negative value indicates severity behavior. The average infit and outfit values of the raters are close to one. This shows that the model-data fit is good.

It is seen that there are two different models of the rater facet population and sample. If the model includes all possible components of the facet, the "model population" should be interpreted according to the values in the "model sample" row (Linacre, 2014). Accordingly, the values in the "model sample" row were interpreted. It is seen that the discrimination rate (6.12) and reliability index (0.96) are high. The reliability index value calculated for the rater facet shows a reliable difference (Haiyang, 2010). This shows that raters exhibit differentiated severity/leniency behaviors. When Table 1 is examined, it is seen that there are fixed effects and normal Chi-square values for the rater facet. The "normal Chi-square" value should be used to examine whether the facet components represent a randomly selected sample from a normally distributed population, and the "fixed-effect Chi-square" value should be used to examine whether there is a difference between the facet components after allowing for measurement error (Linacre, 2014). Accordingly, the fixed-effect Chi-square value was used to examine whether there was a significant difference in terms of the raters' severity and leniency behaviors. The Chi-square values of the rater facet before rater training were statistically significant $\chi^2(sd) = 433.1 (24), p=0.00 < 0.01$. This shows that the raters exhibited differentiated behaviors (severity/leniency). After having determined that the raters exhibited differentiated behaviors at the group level, individual student statistics were examined. While examining the raters' behaviors on a student basis, the *t* value was used. After comparing the obtained *t* value with the critical *t* value in the *t* distribution table, its statistical significance was determined. *t* value was obtained by dividing the difference between the logit value of the rater and the logit mean of all raters by the standard error. The degrees of freedom for the 25 raters before rater training was 24. At a 0.05 level of significance for 24 degrees of freedom, *t* critical was found to be 2.064. The distribution of *t* values for pre-test scores is given in Figure 2.

Figure 2. Distribution of *t* values for pre-test scores.



When Figure 2 is examined, it is seen that 16 (64.00%) of the 25 raters exhibited severity or leniency behavior before rater training. While nine of these raters (36.00%) displayed leniency behavior, seven of them (28.00%) displayed severity behavior. Rater and student interactions were examined to determine differentiated rater severity and leniency at the group level, rater bias in the rater group. Since the Chi-square statistic result of the rater group was significant $\chi^2(sd) = 535.2 (250)$, $p = 0.00 < 0.01$, it was determined that there was a group-level bias effect among the raters. After determining the bias effect at the group level, student-based statistical indicators were examined. In the many-facet Rasch model, a t value outside the ± 2 range indicates significance, that is, rater bias (Linacre, 2023, p. 190). Significant interactions for the pre-test are given in Table 2.

Table 2. Pre-test significant rater-student interactions.

Rater	Student	Observed Score	Expected Score	Bias (Logit)	Standard Error	t
R1	S8	11.00	16.44	-1.05	0.52	-2.02
R2	S7	12.00	18.42	-1.12	0.48	-2.34
R2	S9	25.00	18.67	1.32	0.59	2.25
R4	S9	12.00	21.11	-1.56	0.48	-3.27
R4	S5	10.00	17.00	-1.45	0.59	-2.44
R4	S2	21.00	14.87	0.99	0.42	2.35
R5	S4	12.00	19.81	-1.34	0.48	-2.81
R5	S5	7.00	14.42	-2.86	1.42	-2.02
R5	S7	24.00	18.31	1.07	0.52	2.08
R6	S7	10.00	15.82	-1.26	0.59	-2.13
R6	S3	24.00	17.98	1.12	0.52	2.18
R6	S4	25.00	17.34	1.53	0.59	2.61
R8	S6	12.00	17.80	-1.02	0.48	-2.14
R8	S5	28.00	15.03	3.69	1.41	2.61
R9	S3	20.00	24.17	-0.85	0.41	-2.09
R10	S8	17.00	11.14	1.10	0.40	2.79
R11	S9	26.00	18.67	1.73	0.71	2.44
R11	S6	26.00	17.25	1.95	0.71	2.75
R12	S5	9.00	7.45	1.51	0.72	2.10
R13	S4	15.00	20.84	-0.94	0.41	-2.29
R13	S5	9.00	15.45	-1.63	0.72	-2.26
R13	S8	25.00	13.75	2.12	0.59	3.62
R14	S7	27.00	19.09	2.36	1.00	2.36
R15	S4	28.00	20.23	2.86	1.41	2.03
R17	S10	19.00	12.81	1.03	0.40	2.59
R17	S9	25.00	15.49	1.82	0.59	3.11
R18	S8	23.00	16.80	1.07	0.47	2.27
R19	S10	7.00	14.99	-2.96	1.42	-2.08
R19	S4	26.00	19.27	1.63	0.71	2.30
R19	S6	23.00	16.58	1.10	0.47	2.34
R19	S5	25.00	13.92	2.09	0.59	3.56
R20	S2	17.00	11.96	0.90	0.40	2.29
R21	S3	20.00	24.53	-0.95	0.41	-2.35
R22	S7	18.00	23.14	-0.91	0.39	-2.31
R22	S3	20.00	24.46	-0.93	0.41	-2.29
R23	S9	28.00	20.08	2.88	1.41	2.04
R24	S5	25.00	27.59	-2.00	0.59	-3.42
R25	S2	14.00	7.69	2.46	0.42	5.81

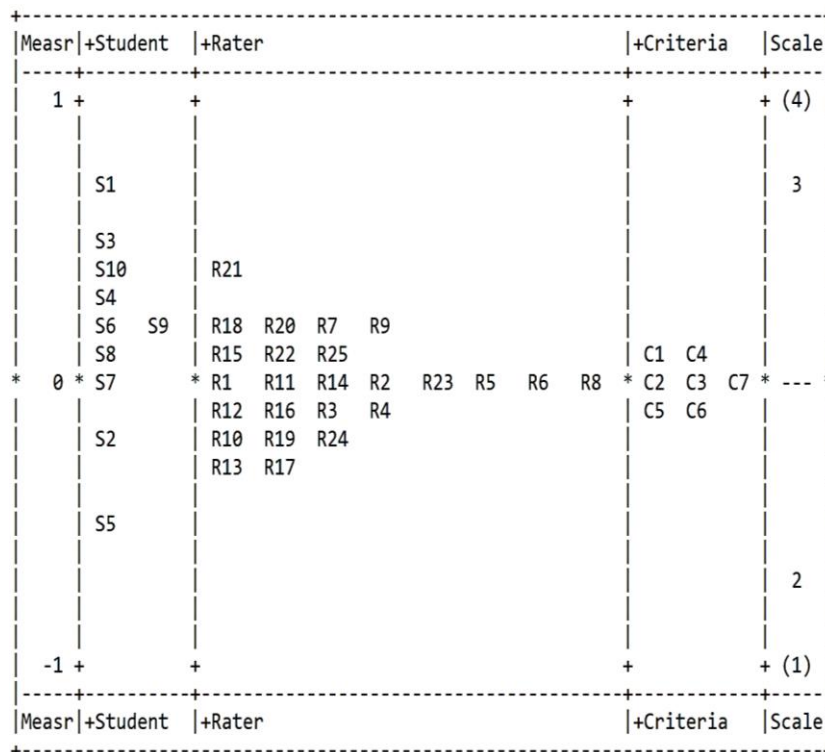
Chi-square = 535.2, $sd = 250$, $p = 0.00$

When Table 2 is analyzed, 38 out of 250 possible interactions between rater and student facets (15.20%) were found to be statistically significant. This indicates that the raters rated some students severely and some students leniently.

3.2. Findings After Rater Training (Post-Test)

The post-test calibration map of peer scores, the rater facet measurement report, rater severity and leniency, and biased interactions measured after rater training within the scope of the study are given below.

Figure 3. Calibration map of peer scores after rater training.



When the calibration map of peer scores after rater training in Figure 3 is examined, it is seen that the student with the highest ability level is S1, the student with the lowest ability level is S5, the most lenient rater is R21, the most severe raters are R13 and R17, the easiest criteria are C1 and C4, and the most difficult criteria are C5 and C6.

The measurement reports of the rater facet for a detailed examination of post-test rater behaviors are given in Table 3. Table 3 shows the observed and adjusted means, logit values, standard error of logit values, concordance, and non-concordance values of the raters after rater training. A positive value in the logit values of the raters indicates leniency, while a negative value indicates severity behavior. The average of the infit and outfit values of the raters is 1.00. This shows that the model-data fit is good. The logit measures of the raters vary between -0.29 and 0.36 and the difference is 0.65. The discrimination rate (0.54) and reliability (0.23) are low. The reliability value calculated for the rater facet shows a reliable difference (Haiyang, 2010). This shows that the raters have similar behaviors. After rater training, the fixed-effect Chi-square values of the rater facet were not statistically significant ($\chi^2(sd) = 32.0(24), p=0.13>0.01$). This is an indication that the raters do not have severity or leniency behavior at the group level. After determining that raters exhibited similar behaviors at the group level, individual statistics on a student basis were examined. The *t* value was used when examining the raters' behaviors on a student basis. After comparing the obtained *t* value with the critical *t* value in the *t* distribution table, its statistical significance was determined. *t* value was obtained by dividing the difference between the logit value of the rater and the logit mean of all raters by the standard

error. The degree of freedom was 24 for the 25 raters after rater training. At a 0.05 level of significance for 24 degrees of freedom, t critical was found to be 2.064. The distribution of t -values for the post-test scores is given in Figure 4.

Table 3. Rater facet post-test measurements measurement report.

Rater	Observed Average	Fair Average	Logit	Model S.E.	Infit	Outfit
R1	2.63	2.63	0.00	0.14	1.07	1.06
R2	2.61	2.62	-0.02	0.14	0.99	0.99
R3	2.54	2.54	-0.11	0.14	1.15	1.16
R4	2.53	2.53	-0.13	0.14	0.88	0.88
R5	2.64	2.64	0.02	0.14	1.35	1.34
R6	2.64	2.64	0.02	0.14	0.82	0.82
R7	2.74	2.75	0.16	0.14	0.51	0.50
R8	2.64	2.64	0.02	0.14	1.36	1.36
R9	2.76	2.76	0.18	0.14	0.79	0.80
R10	2.47	2.47	-0.21	0.14	0.72	0.73
R11	2.60	2.60	-0.04	0.14	1.54	1.55
R12	2.57	2.57	-0.07	0.14	0.92	0.92
R13	2.43	2.43	-0.27	0.14	0.92	0.92
R14	2.61	2.62	-0.02	0.14	1.12	1.11
R15	2.70	2.70	0.10	0.14	0.74	0.74
R16	2.57	2.57	-0.07	0.14	0.99	0.99
R17	2.41	2.41	-0.29	0.14	1.04	1.04
R18	2.76	2.76	0.18	0.14	0.68	0.68
R19	2.47	2.47	-0.21	0.14	1.61	1.58
R20	2.80	2.81	0.24	0.14	1.10	1.11
R21	2.89	2.89	0.36	0.14	0.67	0.68
R22	2.73	2.73	0.14	0.14	0.95	0.95
R23	2.64	2.64	0.02	0.14	1.18	1.19
R24	2.51	2.51	-0.15	0.14	1.22	1.21
R25	2.71	2.72	0.12	0.14	0.61	0.62
Mean	2.63	2.63	0.00	0.14	1.00	1.00
S (population)	0.12	0.12	0.16	0.00	0.28	0.27
S (sample)	0.12	0.12	0.16	0.00	0.28	0.28
Model, Population	RMSE= 0.14		Adj. S.D.= 0.08	Separation= 0.54		
	Strata= 1.06		Reliability= 0.23			
Model, Sample	RMSE= 0.14		Adj. S.D.= 0.08	Separation= 0.59		
	Strata = 1.12		Reliability = 0.26			
Model, Chi-square (Fixed Effect):	32.0	df= 24	p= 0.13			
Model, Chi-square (Normal):	13.8	df= 23	p=0.93			

Figure 4. Distribution of *t* values for post-test scores.

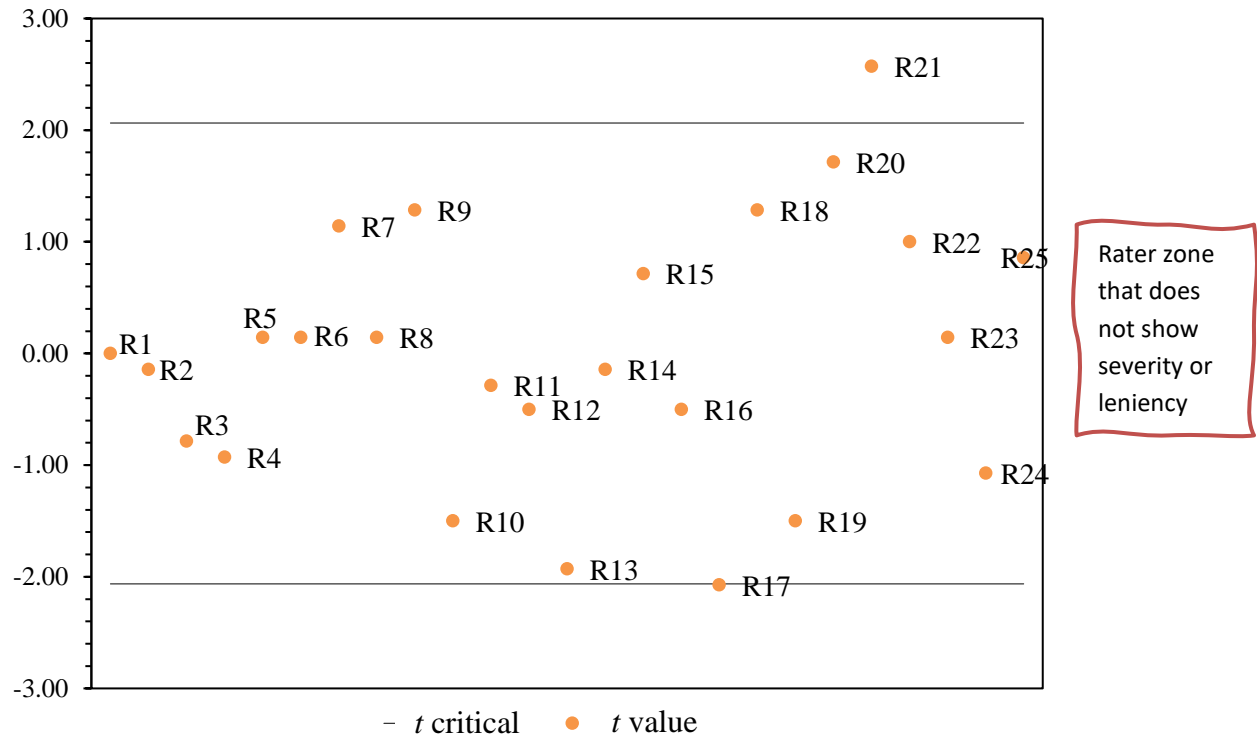


Figure 4 shows that two of the 25 raters (8.00%) exhibited severity or leniency behavior after rater training. One of these raters (4.00%) exhibited leniency behavior, and the other (4.00%) exhibited severity behavior.

The pre-test and post-test statistics of the raters were compared to examine whether there was a statistical difference between rater severity and leniency. For this purpose, *t* statistics, which are indicators of the strictness and generosity of the raters, were compared. Pre-test *t* statistics were compared to post-test *t* statistics as it is an indicator of rater severity/leniency. However, to better observe the impact of rater training, the pre-test *t* statistical value was differentiated from the raters’ post-test *t* statistics. The Mann-Whitney U test results for *t* statistics of pre-test and post-test data are given in Table 4.

Table 4. Mann Whitney U test results of pre-test and post-test *t* statistics.

Test	N	Mean Rank	Sum of Ranks	U	<i>p</i>
Pre-test	25	20.06	501.50	176.50	0.008
Post-test	25	30.94	773.50		
Total	50				

When Table 4 is examined, a statistically significant difference is seen in the raters’ pre-test and post-test severity and leniency behaviors ($U=176.50; p=0.008<0.05$), indicating a statistical difference in rater severity/leniency before and after rater training. It can also be said that this difference is in favor of the post-test when considering the decrease in rater effect after rater training.

Rating and student interaction were studied to determine whether the rater had exclusive behavior at the group level. As a result of Chi-square statistics ($\chi^2(sd) = 389.8 (250), p= 0.00 < 0.01$), it was determined that there was a significant group-level bias effect in the rater group. Student-based statistical indicators were studied after the group-level isolation effect was identified. In the many-facet Rasch model, *t* value outside the ± 2 range indicates significance, punctuation (Linacre, 2023, p. 190). Post-test significant interactions are shown in Table 5.

Table 5. Post-test significant rater-student interactions

Rater	Student	Observed Score	Expected Score	Bias (Logit)	Standard Error	<i>t</i>
R1	S10	12.00	19.35	-1.52	0.52	-2.95
R1	S1	26.00	21.03	1.52	0.74	2.07
R3	S1	14.00	20.47	-1.27	0.46	-2.73
R3	S3	14.00	19.44	-1.06	0.46	-2.29
R4	S10	27.00	18.64	2.74	1.02	2.68
R5	S10	25.00	19.45	1.39	0.62	2.25
R8	S3	14.00	20.13	-1.20	0.46	-2.58
R8	S4	14.00	19.37	-1.05	0.46	-2.26
R8	S9	25.00	18.84	1.51	0.62	2.44
R8	S5	21.00	15.19	1.13	0.46	2.47
R10	S1	26.00	19.99	1.74	0.74	2.35
R11	S10	13.00	19.14	-1.23	0.49	-2.54
R11	S9	25.00	18.54	1.57	0.62	2.53
R12	S8	23.00	17.81	1.09	0.51	2.14
R12	S3	27.00	19.63	2.54	1.02	2.49
R13	S8	22.00	16.79	1.04	0.48	2.17
R15	S7	13.00	18.05	-1.03	0.49	-2.11
R15	S4	26.00	19.77	1.78	0.74	2.41
R16	S5	10.00	14.71	-1.26	0.63	-2.01
R16	S4	24.00	18.86	1.17	0.55	2.12
R16	S3	25.00	19.63	1.36	0.62	2.19
R18	S10	15.00	20.24	-1.01	0.45	-2.25
R19	S6	8.00	17.26	-2.98	1.03	-2.90
R19	S7	10.00	16.42	-1.60	0.63	-2.55
R19	S10	28.00	18.23	3.53	1.43	2.47
R19	S4	26.00	18.15	2.09	0.74	2.84
R20	S3	14.00	21.18	-1.42	0.46	-3.05
R20	S2	23.00	17.57	1.13	0.51	2.23
R22	S2	10.00	17.05	-1.72	0.63	-2.74
R22	S10	26.00	20.04	1.73	0.74	2.34
R23	S10	13.00	19.45	-1.29	0.49	-2.66

Chi-square= 389.8 *sd*= 250 *p*= 0.00

When Table 5 is examined, it is seen that 31 of the possible 250 interactions (12.40%) between the rater and the student facets were statistically significant. This shows that the raters scored some students with severe scores while others with lenient scores.

4. DISCUSSION and CONCLUSION

This study was conducted to determine the effect of rater training, which is one of the methods used to determine and reduce or eliminate rater effect in peer assessment. The many-facet Rasch model was used to determine the rater effect in this study. Pre-test severity and leniency behaviors of the rater group were examined, and as a result, group-level severity and leniency behaviors were observed in the rater group. After the analysis of severity and leniency behaviors at the group level, individual statistics on a student basis were examined. While 16 (64%) of the 25 raters in the rater group were found to be severe or lenient, nine (36.00%) of them were found to have leniency behavior and seven (28.00%) to have severity behavior. Pre-test differentiated rater severity and leniency behaviors at the group level were also included. After the analysis of group-level statistics, the student-level statistics were analyzed. As a result of

the analysis, 38 (15.20%) of the 250 possible interactions between student and rater facets were found to be statistically significant. While 16 of the significant interactions were differentiated rater severity, 22 of them were differentiated rater leniency. These findings are consistent with the studies conducted by Esfandiari and Myford (2013); Farrokhi et al. (2012), Engelhard (1994), Farrokhi and Esfandiari (2011), Karakaya (2015), Şata et al. (2020).

When the post-test severity and leniency of the rater group were examined, it was observed that there was no severity or leniency behavior at the group level. After the analysis of group-level statistics, student-level statistics were analyzed. As a result of the analysis, it was found that two (8.00%) of the 25 raters had severity or leniency behavior: one (4.00%) had rater leniency behavior, and one (4.00%) had rater severity behavior. This may indicate that the two raters may have similar behavior to the pretest.

In addition, a statistically significant difference was found between the raters' pre-test and post-test rater severity and leniency behaviors. This is an indication that rater training was effective in reducing the severity and leniency behaviors of the raters. It was observed that differentiated rater severity and leniency behaviors at the group level continued after rater training. However, only 31 (12.40%) of the 250 possible interactions between the student and rater facets after rater training were found to be statistically significant. While 14 of the significant interactions were differentiated rater severity, 17 were differentiated rater leniency.

Although a decrease in rater effect could be observed after rater training, it did not disappear completely. Many studies investigating the effect of rater training on rater behavior in peer assessment report that rater effect will not change even with feedback or that it will reduce rater behaviors to a certain extent (Berg, 1999; Elder et al., 2005; Knoch, 2011; Knoch et al., 2007; Loignon et al., 2017; Lumley & McNamara, 1995; Lunt et al., 1994; O'Sullivan & Rignall, 2007; Patri, 2002; Wigglesworth, 1993). These studies support the results of this research.

The study sought to explain the possible reasons why rater behaviors did not disappear completely. There are several ways of reducing differential rating inclination and leniency behavior. The first of these methods is to give feedback and rigorous training to the rater. In the study, students did not receive any feedback after rating. Immediate feedback after rating could help raters be more objective when evaluating peers. Knoch (2011) also noted that it would be useful for feedback to raters to be long-term. The lack of feedback in this study may be a cause of bias. There is no standard period in the literature for how long rater training should be given. In this study, students received a total of eight hours of rater training. Giving rater training for an extended period of time may increase the effectiveness of rater training. During rater training, students were given two samples for each criterion. Increasing the number of samples can help students better internalize the criteria. Students (25 students) had limited time to evaluate their peers. This may have caused the raters to misrate some criteria. If the students had had enough time, their scores could have been more objective. Another way could be one-on-one teaching without rater training (Saito, 2008).

The examples used in teaching can make it easier to internalize criteria. The lack of feedback to students and the limited number of samples may have decreased the effect of rater training on rater behavior. Moreover, the task selected for the purpose of this study was persuasive writing. The fact that this type of writing is not included in the Turkish course curriculum may be the reason why rater behaviors have not disappeared.

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(s). **Ethics Committee Number:** Gazi University, 22.04.2022-345966.

Contribution of Authors

Each author has made an equal contribution to the research.

OrcidNazira Tursynbayeva  <https://orcid.org/0000-0002-2165-3276>Umur Öç  <https://orcid.org/0000-0002-1269-1115>İsmail Karakaya  <https://orcid.org/0000-0003-4308-6919>**REFERENCES**

- Alicı, D. (2010). Öğrenci Performansının Değerlendirilmesinde Kullanılan Diğer Ölçme Araç ve Yöntemleri [Other Measurement Tools and Methods Used in the Evaluation of Student Performance (pp. 127-168), Measurement and Evaluation in Education]. Ankara: Pegem Akademi Yayıncılık
- Andrade, H. G. (2005). Teaching With Rubrics: The Good, the Bad, and the Ugly. *College Teaching*, 53(1), 27-31. <https://doi.org/10.3200/CTCH.53.1.27-31>
- Berg, E.C. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*, 8(3), 215-241. [https://doi.org/http://doi.org/10.1016/S1060-3743\(99\)80115-5](https://doi.org/http://doi.org/10.1016/S1060-3743(99)80115-5)
- Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education*, 5(1), 1460901. <https://doi.org/10.1080/2331186X.2018.1460901>
- Boud, D., Cohen, R., & Sampson, J. (1999). Peer Learning and Assessment. *Assessment & Evaluation in Higher Education*, 24(4), 413-426. <https://doi.org/10.1080/0260293990240405>
- Bushell, G. (2006). Moderation of peer assessment in group projects. *Assessment & Evaluation in Higher Education*, 31(1), 91-108. <https://doi.org/10.1080/02602930500262395>
- Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı [Factor Analysis: Basic Concepts and its Use in Scale Development]. *Eğitim Yönetimi: Teori ve Uygulama*, 32(32), 470-483
- Congdon, P.J., & MeQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, 37(2), 163-178. <https://doi.org/https://doi.org/10.1111/j.1745-3984.2000.tb01081.x>
- Çeçen, M. (2011). Türkçe Öğretmenlerinin Seviye Belirleme Sınavı ve Türkçe Sorularına İlişkin Görüşleri [Turkish Language Teachers' Views About Level Determination Exam and Turkish Lesson Questions]. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* 8(15), 201-212. <https://dergipark.org.tr/en/pub/mkusbed/issue/19555/208689>
- Çokluk, Ö., Şekercioglu, G., & Büyüköztürk, Ş. (2021). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları [Multivariate Statistical SPSS and LISREL Applications for Social Sciences]* (6 ed.). Pegem Akademi Yayıncılık <https://doi.org/10.14527/9786055885670>
- Donnon, T., McIlwrick, J., & Woloschuk, W. (2013). Investigating the Reliability and Validity of Self and Peer Assessment to Measure Medical Students' Professional Competencies. *Creative Education*, 4(6), Article 32932. <https://doi.org/10.4236/ce.2013.46A005>
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.
- Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does It Work? *Language Assessment Quarterly*, 2(3), 175-196. https://doi.org/10.1207/s15434311laq0203_1
- Ellington, H., Earl, S., & Cowan, J. (1997). Making effective use of peer and self assessment. *Innovations in Education and Training International*, 32, 175-178.
- Engelhard, G. (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, 31(2), 93-112. <https://doi.org/https://doi.org/10.1111/j.1745-3984.1994.tb00436.x>
- Ertürk, S. (1979). Program development in education (3rd Edition). Yelkentepe Publications.

- Esfandiari, R., & Myford, C.M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*, 18(2), 111-131. <https://doi.org/https://doi.org/10.1016/j.asw.2012.12.002>
- Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *International Journal of Language Testing*, 1(1), 1-16.
- Farh, J.-L., Cannella, A.A., & Bedeian, A.G. (1991). The Impact of Purpose on Rating Quality and User Acceptance. *Group & Organization Studies*, 16(4), 367-386. <https://doi.org/10.1177/105960119101600403>
- Farrokhi, F., & Esfandiari, R. (2011). A Many-facet Rasch Model to Detect Halo Effect in Three Types of Raters. *Theory & Practice in Language Studies*, 1(11).
- Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34(1), 79-101.
- Field, A. (2005). Reliability analysis. *Discovering Statistics Using spss. 2nd Edition*, Sage, London.
- Greenan, K., Humphreys, P., & McIlveen, H. (1997). Developing transferable personal skills: part of the graduate toolkit. *Education + Training*, 39(2), 71-78. <https://doi.org/10.1108/00400919710164161>
- Guadagnoli, E., & Velicer, W.F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265-275. <https://doi.org/10.1037/0033-2909.103.2.265>
- Gürten, E., Boztunç Öztürk, N., & Eminoğlu, E. (2019). Investigation of the Reliability of Teacher, Self and Peer Evaluations at Primary School Level Using Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, 10(4), 406-421.
- Haiyang, S. (2010). An application of classical test theory and many facet Rasch measurement in analyzing the reliability of an English test for non-English major graduates. *Chinese Journal of Applied Linguistics*, 33(2), 87-102.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.
- Hauenstein, N.M.A., & McCusker, M.E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment*, 25(3), 253-266. <https://doi.org/https://doi.org/10.1111/ijsa.12177>
- Heslin, P.A. (2005). Conceptualizing and evaluating career success. *Journal of Organizational Behavior*, 26(2), 113-136. <https://doi.org/https://doi.org/10.1002/job.270>
- Hutcheson, G.D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage.
- İşman, A., & Eskicumalı, A. (2003). Eğitimde Planlama ve Değerlendirme [Planning and Evaluation in Education] (4th Edition). *Istanbul: Değişim Yayınları*
- Johnson, C., & Smith, F. (1997). Assessment of a complex peer evaluation instrument for team learning and group processes. *ACCOUNTING EDUCATION-GREENWICH*, 2, 21-40.
- Karakaya, İ. (2015). Comparison of Self Peer and Instructor Assessments in the Portfolio Assessment by Using Many Facet Rasch Model. *Journal of Education and Human Development*, 4(2).
- Keaten, J.A., & Richardson, M.E. (1993). A Field Investigation of Peer Assessment as Part of the Student Group Grading Process.
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior - a longitudinal study. *Language Testing*, 28(2), 179-200. <https://doi.org/10.1177/0265532210384252>

- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26-43. <https://doi.org/https://doi.org/10.1016/j.asw.2007.04.001>
- Kondo, Y. (2010). Examination of Rater Training Effect and Rater Eligibility in L2 Performance Assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, 14(2), 1-23.
- Kubiszyn, T., & Borich, G.D. (2024). *Educational testing and measurement*. John Wiley & Sons.
- Kurudayioğlu, M., Şahin, Ç., & Çelik, G. (2008). Türkiye’de Uygulanan Türk Edebiyatı Programı’ndaki Ölçme ve Değerlendirme Boyutu Uygulamasının Değerlendirilmesi: Bir Durum Çalışması [Evaluation of the Application of Measurement and Evaluation Dimension in Turkish Literature Program Implemented in Turkey: A Case Study]. *Ahi Evran University Kırşehir Eğitim Fakültesi Dergisi*, 9(2), 91-101. <https://dergipark.org.tr/en/pub/kefad/issue/59525/856034>
- Kutlu, Ö., Doğan, C.D., & Karakaya, İ. (2010). Öğrenci başarısının belirlenmesi performans ve portfolyoya dayalı durum belirleme [Determining student achievement based on performance and portfolio assessment]. Ankara: Pegem Akademi Yayıncılık
- Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563-575. <https://doi.org/https://doi.org/10.1111/j.1744-6570.1975.tb01393.x>
- Linacre, J.M. (2014). *A user’s guide to FACETS: Rasch-model computer programs* (Vol. 18). <http://www.winsteps.com/manuals.htm>
- Linacre, J.M. (2023). *Facets computer program for many-facet Rasch measurement*. Winsteps.com.
- Loignon, A.C., Woehr, D.J., Thomas, J.S., Loughry, M.L., Ohland, M.W., & Ferguson, D.M. (2017). Facilitating peer evaluation in team contexts: The impact of frame-of-reference rater training. *Academy of Management Learning & Education*, 16(4), 562-578. <https://doi.org/10.5465/amle.2016.0163>
- Lumley, T., & McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12(1), 54-71. <https://doi.org/10.1177/026553229501200104>
- Lunt, H., Morton, J., & Wigglesworth, G. (1994). Rater behaviour in performance testing: Evaluating the effect of bias feedback. 19th annual congress of Applied Linguistics Association of Australia: University of Melbourne. July,
- Martin, C.C., & Locke, K.D. (2022). What Do Peer Evaluations Represent? A Study of Rater Consensus and Target Personality [Brief Research Report]. *Frontiers in Education*, 7. <https://doi.org/10.3389/educ.2022.746457>
- May, G.L. (2008). The Effect of Rater Training on Reducing Social Style Bias in Peer Evaluation. *Business Communication Quarterly*, 71(3), 297-313. <https://doi.org/10.1177/1080569908321431>
- Myford, C.M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, 4(4), 386-422.
- Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement*, 5(2), 189-227.
- O’Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. *IELTS Collected Papers: Research in speaking and writing assessment*, 446-478.
- Oosterhof, A. (1999). *Developing and using classroom assessments*. ERIC.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19(2), 109-131. <https://doi.org/10.1191/0265532202lt224oa>
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581. <https://doi.org/10.1177/0265532208094276>

- Somervell, H. (1993). Issues in Assessment, Enterprise and Higher Education: the case for self-peer and collaborative assessment. *Assessment & Evaluation in Higher Education*, 18(3), 221-233. <https://doi.org/10.1080/0260293930180306>
- Stiggins, R., & Chappuis, J. (2005). Using Student-Involved Classroom Assessment to Close Achievement Gaps. *Theory Into Practice*, 44(1), 11-18. https://doi.org/10.1207/s15430421tip4401_3
- Şata, M., Karakaya, İ., & Erman Aslanoğlu, A. (2020). Evaluation of University Students' Rating Behaviors in Self and Peer Rating Process via Many Facet Rasch Model [Üniversite Öğrencilerinin Öz ve Akran Puanlama Sürecinde Puanlama Davranışlarının Many Facet Rasch Modeli ile İncelenmesi]. *Eurasian Journal of Educational Research*, 20(89), 25-46. <https://dergipark.org.tr/en/pub/ejer/issue/57497/815802>
- Turgut, M.F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]* (Vol. 2). Ankara: Pegem Akademi Yayıncılık
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-319. <https://doi.org/10.1177/026553229301000306>
- Woolfolk, A.E., Hoy, A.W., Hughes, M., & Walkup, V. (2008). *Psychology in education*. Pearson Education.
- Yaşar, M. (2017). Ölçme ve değerlendirmenin önemi [The importance of measurement and evaluation]. *Pegem Citation Index*, 2-8.
- Yeşilçınar, S., & Şata, M. (2021). Examining Rater Biases of Peer Assessors in Different Assessment Environments. *International Journal of Psychology and Educational Studies*, 8(4), 136-151. <https://dergipark.org.tr/en/pub/pes/issue/65718/1020683>

How many grades of response categories does the commitment to the profession of medicine scale provide the most information?

Murat Tekin¹, Çetin Toraman^{1*}, Ayşen Melek Aytuğ Koşan¹

¹Çanakkale Onsekiz Mart University, Faculty of Medicine, Department of Medical Education, Çanakkale, Türkiye

ARTICLE HISTORY

Received: Dec. 04, 2023

Accepted: July 24, 2024

Keywords:

Likert scale,
Response set,
Item response theory,
Medical student.

Abstract: In the present study, we examined the psychometric properties of the data obtained from the Commitment to Profession of Medicine Scale (CPMS) with 4-point, 5-point, 6-point, and 7-point response sets based on Item Response Theory (IRT). A total of 2150 medical students from 16 different universities participated in the study. The participants were divided into four groups consisting of 560, 544, 502, and 544 medical students. The first group (n=560) was assigned four-point, the second group (n=544) five-point, the third group (n=502) six-point, and the fourth group (n=544) seven-point Likert forms. We used R statistical software to analyze the data. The results of item calibrations conducted with the Graded Response Model (GRM) were analyzed. The results show that the eigenvalue increased from 4-point to 7-point. Similarly, the explained variance percentage and the scale's reliability increased gradually from 4-point to 7-point. The explained variance, reliability level, and eigenvalue were very close in the 5-point and 6-point forms.

1. INTRODUCTION

Scales are used to collect data in many scientific fields. Scales can be configured with the Thurstone scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017; Torgerson, 1958), Guttman scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Lord, 1954; Nunnally & Bernstein, 1994; Price, 2017), and the Likert scaling technique (Anastasi & Urbina, 1997; Dunn-Rankin et al., 2004; Price, 2017). In Likert-type scales, mainly used to measure thoughts, beliefs, and attitudes, the participants' level of agreement in the statements given is measured by grading with the Likert scaling technique (Anastasi & Urbina, 1997; DeVellis, 2003). Likert scales are very popular in use as they are easy to configure. Likert scales are widely used in social sciences and educational research (Joshi et al., 2015).

When taking the participants' answers, distances between each choice (answer option) are assumed to be equal in Likert scales. This is because Likert (1932) suggests that the "distance between response categories is assumed to be equal". Response set may broadly include five points to a statement: (1) Strongly disagree, (2) Disagree, (3) Neither agree nor disagree, (4) Agree, and (5) Strongly agree (Anastasi & Urbina, 1997). Additionally, they may include six

*CONTACT: Çetin TORAMAN ✉ toramanacademic@gmail.com 📧 Çanakkale Onsekiz Mart University, Faculty of Medicine, Department of Medical Education, Çanakkale, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

points to a statement as well: (1) Disagree very strongly, (2) Disagree strongly, (3) Disagree, (4) Agree, (5) Agree strongly, and (6) Agree very strongly (DeVellis, 2003). Likert types can have response sets with or without the ‘neutral’ option. Although there are grade suggestions such as “neither agree nor disagree” or “equally agree and disagree” for the neutral point, discussions regarding this neutral expression continue (DeVellis, 2003). Although the average scores on the Likert scale were not substantially affected by the inclusion or exclusion of a "Neutral" option, significant variations emerge when combining neighboring categories, such as the proportion of respondents who "Agree or Strongly Agree." This suggests that the presence or absence of a neutral category can lead to considerably different interpretations. Respondents may find the neutral option valuable, and its removal could result in misleading conclusions, especially when analyzing individual items. In contexts where the scale is used for quality enhancement or progress tracking, including a neutral option may provide a more accurate reflection of shifts in perception. (Mariano et al., 2024). It is important to determine whether Likert scales obtain data at the ordinal or interval scale through response categories. While some researchers (Jamieson, 2004; Stevens, 1946; Thomas, 1982) claim that the data obtained from Likert scales are at the ordinal scale level, some others (Norman, 2010) assert that it can be accepted at an interval scale level and parametric analyzes can be used in line with this assumption. Some studies suggest that by increasing the number of grades in the answer set, the obtained data will be normally distributed and set to an interval scale level (Wu & Leung, 2017). Several studies have been conducted on the descriptive statistics of data obtained from Likert scales with varying response categories using 5, 7, and 10-degree response sets. It was determined that the scale mean with 10 response categories tended to be lower than the scale mean with 5 or 7 response categories. The scales offered very similar values in terms of skewness and kurtosis (Dawes, 2008). In another study conducted with Likert scales with different response categories (4, 5, 6, and 11 response categories), no major differences could be determined between the mean, standard deviation, item correlations, Cronbach Alpha value, and factor loadings of the data obtained. The skewness and kurtosis of the data obtained from the scale with the most response categories (11 degrees) decreased and approached normal distribution (Leung, 2011). In another study, the data obtained from the Likert scale, prepared in different forms as 5, 7, 9, and 11 response categories, were compared in terms of mean, standard deviation, skewness, and kurtosis. The increase in response categories caused the mean to decrease. According to the skewness value, the closest scale to the normal distribution is the 5-degree scale. According to the kurtosis value, the scale closest to the normal distribution is the 11-category scale (Bora, 2013).

Likert-type scales have been the subject of extensive research studies on:

- The effect of the number of categories in the response set on the alpha coefficient (Aiken, 1983; Chang, 1994; Leung, 2011; Wong et al., 1993),
- Its effect on test-retest reliability level (Preston & Colman, 2000),
- How many grades an answer set should have (Champney & Marshall, 1939),
- How the number of grades in the response set affects the arithmetic means and distribution measures (standard deviation, kurtosis, skewness) of the data obtained (Bora, 2013; Dawes, 2008; Leung, 2011),
- Its effects on the normal distribution (Leung, 2011),
- Participants’ perceptions of variables in the answer set (Adelson & McCoach, 2010),
- How the number of grades in a response set affects item parameters based on item response theory (IRT) (Aybek & Toraman, 2022; Wakita et al., 2012).

In summary, agreement categories of relevant Likert model scales were examined based on reliability, covariance matrices, descriptive statistics, the ability to distinguish the neutral option in the response set, and the effect on factor loads in terms of classical test theory (CTT). Aybek and Toraman, (2022) and Wakita et al., (2012) examine the effect of the number of grades in

IRT on the item's functioning with its options. This research contributes to IRT-based studies by analyzing how scales with 4-point ("Strongly Disagree", "Disagree", "Agree", "Strongly Agree"), 5-point ("Strongly Disagree", "Disagree", "Undecided", "Agree", "Strongly Agree"), 6-point ("Strongly Disagree", "Disagree", "Somewhat Disagree", "Somewhat Agree", "Agree", "Strongly Agree"), and 7-point ("Strongly Disagree", "Disagree", "Somewhat Disagree", "Neither Agree nor Disagree", "Somewhat Agree", "Agree", "Strongly Agree") response sets work. The findings indicate that the number of scale points significantly impacts the perceived psychological distance between options, particularly for seven-point scales. In this study, the "Commitment to Profession of Medicine Scale (CPMS)" comprising 4-point, 5-point, 6-point, and 7-point response sets by Aytug Kosan and Toraman (2020), was used. The researchers who developed the CPMS developed this scale with five response categories (strongly disagree, disagree, partially agree, agree, and completely agree). This study examines the psychometric properties of the data obtained from the scale with 4-point, 5-point, 6-point, and 7-point response sets based on IRT.

2. METHOD

2.1. Participants

A total of 2150 medical students from 16 different universities participated in the study. Participants were divided into 4 groups with 560, 544, 502, and 544 medical students. In this study, the CPMS was used as the data collection tool; and the groups were given 4-point, 5-point, 6-point, and 7-point Likert forms of CPMS, respectively. The first group (n=560) was assigned 4-point, the second group (n=544) 5-point, the third group (n=502) 6-point, and the fourth group (n=544) 7-point Likert forms. The distribution of the participants by gender, study year, and university-type variables is given in [Table 1](#).

Table 1. Descriptive statistics on participants' sex, study year, and university type variables.

	Variable	4-point	5-point	6-point	7-point
Sex	Female	300	285	294	280
	Male	260	259	208	264
Year	Preparatory	16	---	---	---
	Year 1	114	226	64	104
	Year 2	190	131	53	81
	Year 3	35	54	89	31
	Year 4	28	36	14	69
	Year 5	140	72	188	30
	Year 6	37	25	94	229
University	State	430	402	462	406
	Foundation (Private)	130	142	40	138

2.2. Measurement Tool

The data were obtained using the Commitment to Profession of Medicine Scale (CPMS), which scale was developed by Aytug Kosan and Toraman (2020) and comprised nine items. The original version of the scale has a 5-point Likert structure (strongly agree, agree, partly agree, disagree, strongly disagree). Within the scope of this research, 4-point, 5-point, 6-point, and 7-point forms of the scale were created and applied to four different groups. Aytug Kosan and Toraman (2020) have reported their scale's validity and reliability evidence through exploratory factor analysis (EFA), confirmatory factor analysis (CFA), IRT, Cronbach Alpha, and marginal reliability coefficient. As a result of factor analysis, the structure of the scale was set as 9 items and a single factor.

2.3. Procedure

- The ethics committee approval was obtained for the study.
- This study was approved by the relevant medical faculty rectors and faculty deans.
- The medical faculties to which the CPMS with a 4-point, 5-point, 6-point, and 7-point Likert answer set would be sent was determined.
- Through the faculty deans, the information about the purpose of the research and how the data collection process would be was shared with the students.
- The scales were delivered online to the students who voluntarily agreed to participate and answer the scales.
- The data were taken from the online environment, transferred to statistical software, and analyzed.

2.4. Data Analysis

Data collected from the participants were analyzed on R 4.1.0 (R Core Team, 2021) using *mirt* 1.35.1 (Chalmers, 2012) and *psych* 2.1.6 (Revelle, 2021) packages. In addition, *MVN* 5.9 (Korkmaz et al., 2014) package was used to determine whether the data showed a multivariate normal distribution. In the data analysis, the tested topics, respectively, are:

- Multivariate normality (Henze-Zirkler Test),
- Unidimensionality can be determined by correlation matrix examination or factor analysis while unidimensionality can be determined using factor analytical techniques (Exploratory Factor Analysis [EFA], Principal Axis Factoring [PAF], Eigenvalue),
- The average variance extracted (AVE) and composite reliability (CR) of the scale were investigated for convergent validity. For these two specified values, $AVE \geq 0.5$ and $CR \geq 0.7$ are required (Fornell & Larcker, 1981)
- Local independence is a fundamental assumption in item response theory (IRT) models. This assumption states that the responses to one item are independent of the responses to other items at a specific level of ability. This does not imply the absence of correlation between items across all groups; rather it indicates that the responses to an item are independent at different levels of proficiency. To fulfill the local independence assumption, it is essential to meet the one-dimensionality assumption. In a one-dimensional model, if item responses are not locally independent, it indicates a multidimensionality dependency. While one-dimensionality is considered sufficient to meet the local independence assumption, additional methods are employed to specifically assess local independence. One such method is the Q_3 test proposed by Yen (1984). This test evaluates local independence between pairs of items by calculating the residuals of each individual's item responses, based on the estimated item parameters. Yen (1984) recommends that researchers treat items with a linear correlation coefficient exceeding 0.20 as potential violators of local independence. This revised text emphasizes key concepts, uses more precise terminology, and avoids unnecessary repetition. It also integrates the information smoothly and provides a clearer understanding of the concept of local independence in the context of IRT models.
- Item-model fit evaluated with S_{χ^2} statistic: According to Browne and Cudeck (1993), the fit indicator in the RMSEA values of the S_{χ^2} statistic is considered as 0.05 and below, and according to Hu and Bentler (1999), as 0.06 and below.
- Item-total correlations, internal consistency (Cronbach α), and marginal reliability levels: Hair, et al. (2014), in social sciences, where information is generally less certain, a solution that meets 60% (and sometimes even less) of the total variance is satisfactory. According to Warner (2013), the acceptable limits are between 40% and 70%. While according to Nunnally and Bernstein (1994), sufficient reliability should be at least 0.70 and above.
- Graded Response Model (GRM): GRM is a ranked response model that assumes the same threshold parameters that define the uniform-ordered categorical response formats category

boundaries. The CPMS structure is also suitable for this modeling. For this reason, modeling was done with GRM.

- Item calibrations made with IRT (GRM): GRM is estimated using marginal maximum likelihood (MML); where the scale is fixed using the latent density function $g(0)$ where the mean and variance are constrained. By convention, $g(0)$ is assumed to be the standard normal density (mean zero and standard deviation one) (Smits et al., 2020). In calibration, one aims to train the item parameters in the IRT model using responses from a sample of the target population. Item calibrations were carried out in accordance with the GRM assumption.
- According to Item Response Theory (IRT), the optimal discrimination parameter ("a" parameter) for an ideal scale item should fall between 0.5 and 2. Research suggests that a discrimination parameter within the range of 0.75 to 2.50 is considered acceptable (Flannery et al., 1995).
- The ideal range for item difficulty levels, as represented by the "b" parameter in Item Response Theory (IRT), is typically considered to be between -1.00 and 1.00, indicating a medium difficulty level (Hambleton, 1994). In inability or achievement tests, items with difficulty levels below -1.00 are generally classified as easy, while those with difficulty levels above 1.00 are considered difficult.
- Option Characteristic Curves (OCC) were examined. OCCs correlate the probability of confirming an item's response options with increasing levels of the trait being measured (Sodano et al., 2014).

3. RESULTS

We conducted a multivariate normal distribution test on CPMS datasets containing Likert answer sets with 4-point, 5-point, 6-point, and 7-point scales. The results did not demonstrate multivariate normal distribution. However, factor analysis revealed that the scales exhibit a one-dimensional structure. Table 2 presents the eigenvalues obtained from Exploratory Factor Analysis (EFA), along with the corresponding variance explained, Cronbach's α , AVE, CR, and marginal reliability coefficients.

Table 2. EFA, explained variance, and reliability coefficients.

	4-point Likert	5-point Likert	6-point Likert	7-point Likert
KMO	0.784	0.877	0.852	0.848
Bartlett's Test of Sphericity	1583.437 ($df=36$, $p<.05$)	2495.955 ($df=36$, $p<.05$)	2409.577 ($df=36$, $p<.05$)	4804.329 ($df=36$, $p<.05$)
Eigenvalues	3.11	4.46	4.38	6.01
Variance explained	35%	50%	49%	67%
Cronbach α	0.81	0.89	0.89	0.95
r_{jx}	0.85	0.90	0.91	0.95
AVE	0.40	0.49	0.49	0.67
CR	0.85	0.89	0.89	0.95

While the eigenvalue was almost identical in the 5-point and 6-point forms, it increased gradually from the 4-point form to the 7-point form. Similarly, the variance explanation and reliability coefficient increased gradually from the 4-point form to the 7-point form. The variance explained and reliability levels in the 5-point and 6-point forms were very close. There are different opinions about how the factor structure obtained should explain the variance of the desired feature. 5-point, 6-point, and 7-point forms achieved the level of variance explanation suggested by the literature. 4-point, 5-point, 6-point and 7-point forms provided reliability at the level suggested by the literature. AVE and CR rates at the level suggested by the literature occurred in forms with 5-point, 6-point, and 7-point response categories.

Yen's Q_3 statistics (Yen, 1993) were used to determine whether the items met the local independence assumption, and local independence was provided in all four forms. At this stage, 0.20 was used as the criterion value for the Q_3 statistic. The item-model fit was examined with the S_{χ^2} statistic. At this stage, the GRM was used as the IRT model. GRM is a polytomous IRT model designed especially for variables accepted as ordinals (Samejima, 2005). The RMSEA values of the S_{χ^2} statistic calibrated according to the GRM and showing the item parameters and item model fit are given in Table 3.

Table 3. Parameter estimation results of CPMS Items.

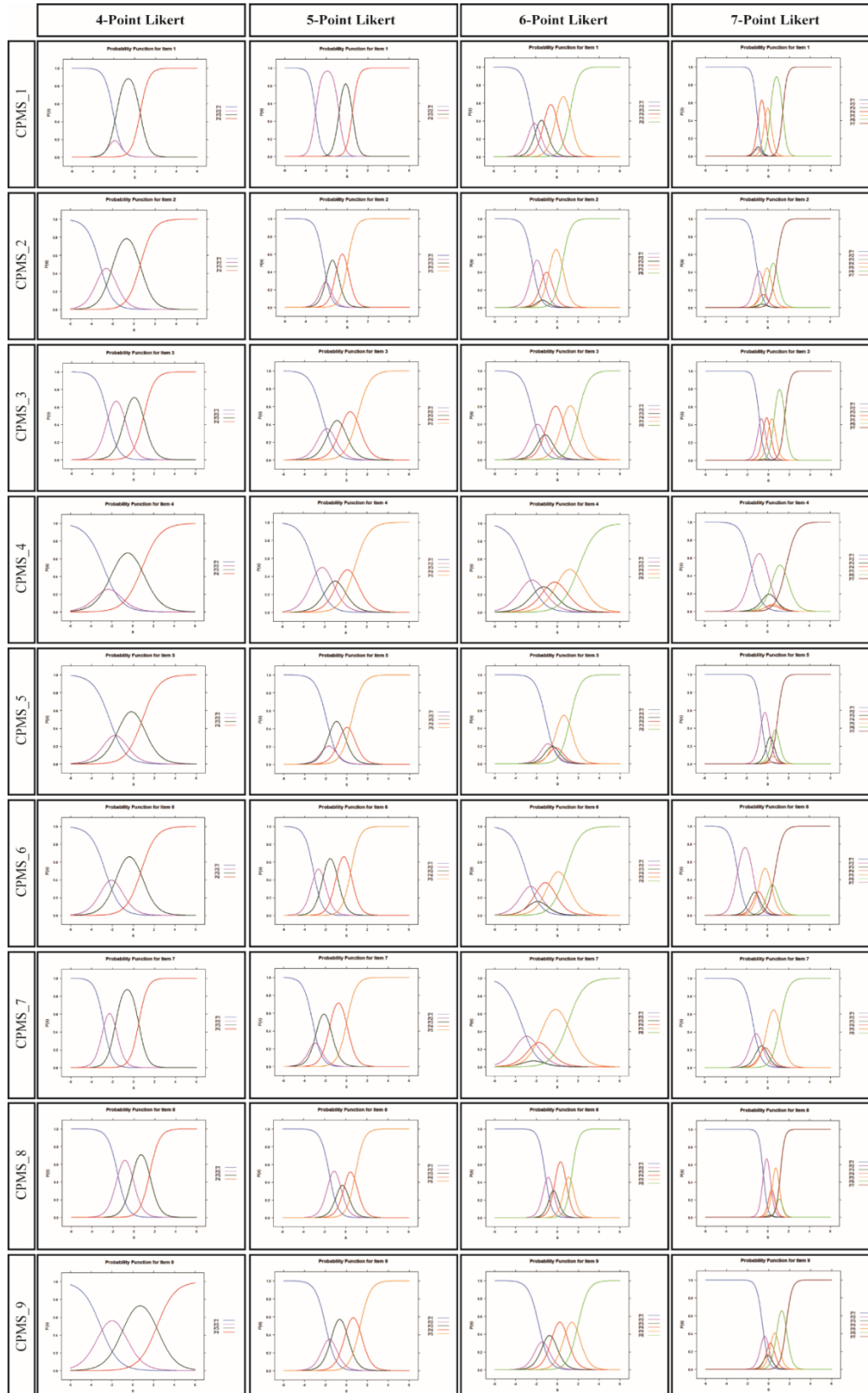
		Items								
		CPMS	CPMS	CPMS	CPMS	CPMS	CPMS	CPMS	CPMS	CPMS
		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
4-point	a	2.45	1.63	1.95	1.16	1.29	1.42	2.45	2.15	1.16
	b ₁	-2.06	-3.21	-2.50	-2.79	-2.27	-2.66	-2.87	-1.56	-3.09
	b ₂	-1.75	-2.00	-0.85	-1.88	-1.24	-1.48	-1.72	-0.14	-0.89
	b ₃	0.53	0.60	0.98	0.90	0.85	0.76	0.49	1.51	2.32
	RMSEA _{Sχ^2}	0.11	0.12	0.12	0.13	0.11	0.15	0.13	0.11	0.12
5-point	a	3.58	2.74	1.72	1.51	1.96	2.38	2.27	2.47	2.06
	b ₁	-3.04	-2.31	-2.31	-2.95	-1.92	-3.16	-3.22	-1.56	-1.98
	b ₂	-0.79	-1.45	-1.45	-1.50	-1.49	-2.19	-2.74	-0.62	-1.27
	b ₃	0.51	-0.34	-0.34	-0.54	-0.42	-0.92	-1.55	-0.01	-0.01
	b ₄	---	1.07	1.07	0.83	0.48	0.42	0.03	0.92	1.32
	RMSEA _{Sχ^2}	0.09	0.07	0.12	0.07	0.10	0.09	0.11	0.08	0.08
6-point	a	2.50	2.62	1.98	1.23	2.11	1.49	1.32	3.18	2.06
	b ₁	-2.43	-2.39	-2.32	-3.03	-1.13	-3.02	-3.47	-1.18	-1.78
	b ₂	-1.80	-1.48	-1.47	-1.79	-0.69	-2.11	-2.38	-0.56	-1.15
	b ₃	-1.11	-1.35	-0.89	-0.83	-0.31	-1.68	-2.17	-0.17	-0.36
	b ₄	-0.04	-0.71	0.52	0.32	0.03	-0.64	-1.32	0.77	0.81
	b ₅	1.26	0.49	1.95	2.03	1.19	0.80	1.02	1.39	1.96
	RMSEA _{Sχ^2}	0.13	0.11	0.09	0.13	0.11	0.11	0.10	0.11	0.11
7-point	a	4.81	3.42	4.27	2.05	3.84	2.76	2.57	4.92	3.69
	b ₁	-1.09	-1.10	-0.90	-1.54	-0.61	-2.85	-1.40	-0.45	-0.58
	b ₂	-1.03	-0.59	-0.42	-0.04	0.08	-1.40	-0.77	0.21	-0.16
	b ₃	-0.94	-0.54	-0.41	0.35	0.40	-1.02	-0.38	0.23	0.02
	b ₄	-0.32	-0.37	0.08	0.48	0.42	-0.63	-0.03	0.50	0.35
	b ₅	0.19	0.20	0.55	0.64	0.51	0.22	1.17	1.02	0.82
	b ₆	1.37	0.84	1.56	1.76	0.93	0.74	---	1.19	1.67
	RMSEA _{Sχ^2}	0.13	0.13	0.17	0.15	0.13	0.12	0.15	0.15	0.15

In the analysis of CPMS data sets applied with 4-point, 5-point, 6-point, and 7-point Likert response sets, the RMSEA values of the S_{χ^2} statistic varied between 0.07 and 0.17. The closest fit to the values determined by the literature was obtained in the 5-point Likert form.

There were mathematical differences in the item discrimination "a" parameters of the four forms. It was determined mathematically that the scale items in 4-point and 6-point forms approached the ideal level of discrimination. The increase in the number of grades in the Likert response set of the scale can be said to increase discrimination. In the context of using the Generalized Rating Scale Model (GRM) as an Item Response Theory (IRT) model, the 'b' parameters representing item confirmation difficulty indicate the level of theta at which the likelihood of selecting categories 2 and 3 equals the likelihood of selecting category 1, and the likelihood of selecting category 3 equals the likelihood of selecting categories 1 and 2. The b parameters increased from the first response category to the last response category for all four forms.

Option Characteristic Curves (OCC), item information function, test information function, and reliability functions were obtained after item calibrations. OCCs were examined to better understand how the number of categories changes the response behavior. The OCCs of 4-point, 5-point, 6-point, and 7-point response categories for all items are given in Figure 1.

Figure 1. OCCs of the items of the CPMS forms administered with a 4-point, 5-point, 6-point, and 7-point Likert answer sets.



When the Option Characteristic Curves (OCCs) are examined, a summary similar to [Table 4](#) can be made.

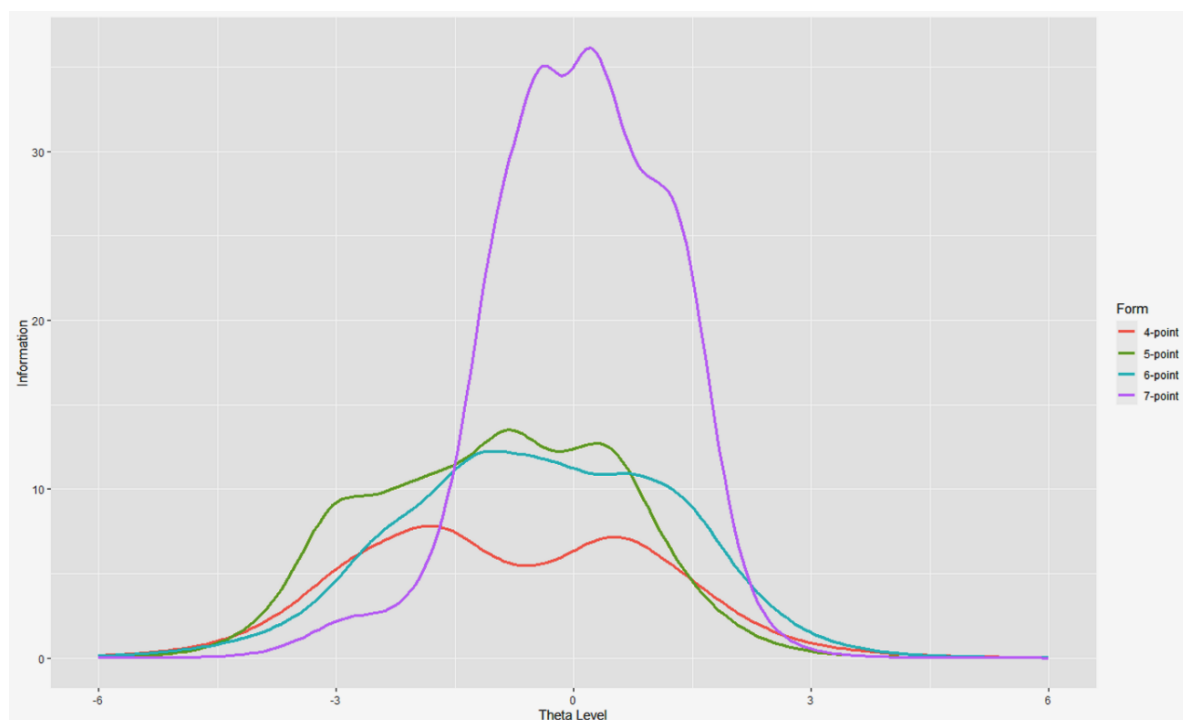
Table 4. Option functioning states of items according to OCC review.

Results	4-point Likert	5-point Likert	6-point Likert	7-point Likert
There are Options that Work Well	Items 2, 3, 7, 8, and 9	Items 4, 6, and 8	Items 1, 3, 4, and 8	Items 3, 6, 7, and 9
There is an Option that Never Works	---	Item 1	---	Item 7
There is an Option that Does Not Differ from Other Options	Items 1, 4, and 5	Items 2, 3, 5, 7, and 9	Items 1, 2, 5, 6, 7, and 9	Items 1, 2, 4, 5, and 8
There are very few Responsive Options	Items 1 and 3	Items 2, 5, and 7	Items 2, 5, 6, and 7	Items 1, 2, 4, 5, and 9

The item options differentiated and worked better in the 4-point Likert form. Additionally, in the 5-point and 7-point Likert forms, there was an item with at least one dysfunctioning option. The number of items with an undifferentiated option from other options was the least in the 4-point Likert form. The number of items with options that received a small response from the participating medical school students was also the least in the 4-point Likert form. As seen in [Tables 2](#) and [3](#), the 4-point Likert form least explained the variance of the scale's measured feature and the item-model fit parameters were not at the level suggested by the literature. However, the 4-point Likert form worked well in identifying the item options and obtaining the participants' responses.

The CPMS forms applied with 4-point, 5-point, 6-point, and 7-point Likert answer sets that gave information with a total of 9 items were examined. The test information functions of the four forms are presented in [Figure 2](#).

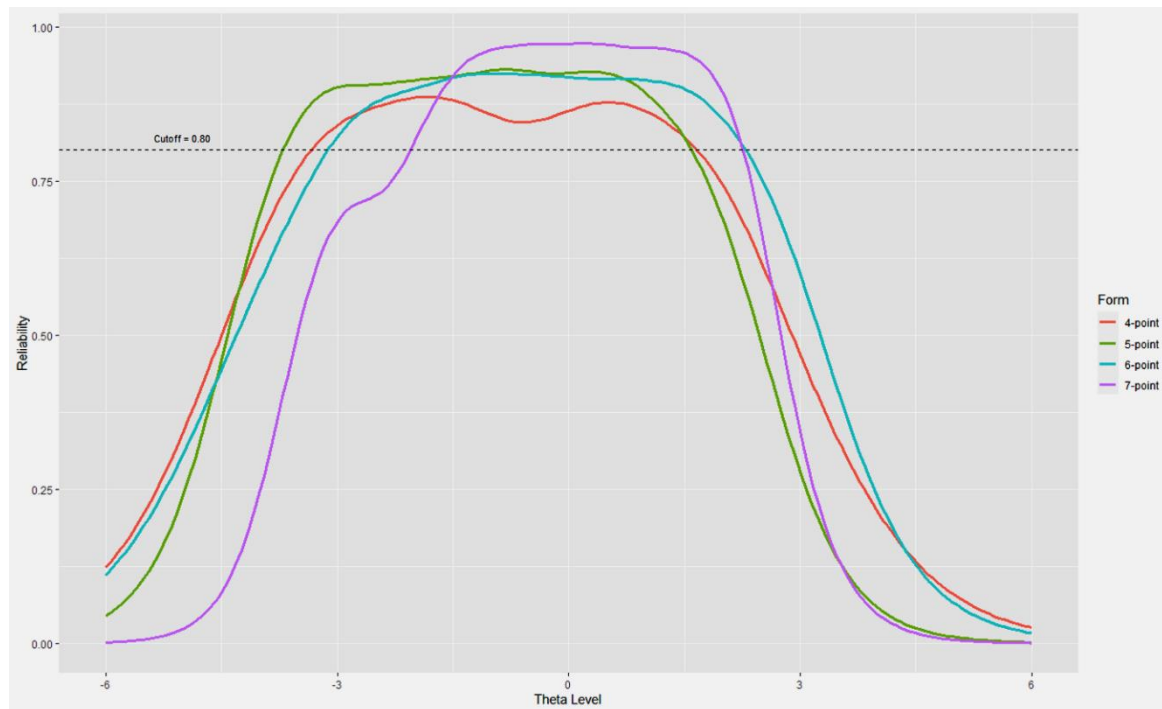
Figure 2. Test information functions of the four forms.



When the test information functions are examined, the form that provides the least information is the 4-point Likert form. Additionally, 5-point and 6-point Likert forms gave similar information. However, 5-point and 6-point Likert forms gave higher information than 4-point

Likert forms and lower than 7-point Likert forms. The most informative form was the 7-point Likert form. The reliability functions obtained for the four forms are presented in [Figure 3](#).

Figure 3. Reliability functions of the four forms.



When the reliability functions were examined, the levels of all four forms exceeded 0.80 and were reliable at a similar level. The form with the highest reliability was the 7-point Likert form, albeit by a small margin. On the other hand, the 4-point, 5-point, and 6-point Likert forms were similar and had higher internal consistency in a slightly wider theta range compared to the 7-point Likert form.

4. DISCUSSION and CONCLUSION

This study investigates the psychometric properties of data collected using a scale with 4-point, 5-point, 6-point, and 7-point response options, employing Item Response Theory (IRT) as the analytical framework. In the research, data obtained with 4-point, 5-point, 6-point, and 7-point response categories forms were analyzed based on IRT. The psychometric evidence obtained pertains to the information presentation levels of the scale items. While the eigenvalue is almost identical in the 5-point and 6-point graded forms, it increases gradually from the 4-point form to the 7-point form. Similarly, the variance disclosure percentage of the scale's measured feature and the scale data's reliability level have increased gradually from the 4-point to the 7-point form. The variance and reliability levels explained in the 5-point and 6-point forms were very close. In the study by Aybek and Toraman (2022), the reliability coefficient of the scale was calculated for the 4-point, 5-point, and 7-point forms. The more categories a form had, the higher reliability values were reached. In addition, researchers could not obtain a multivariate normal distribution in the data set similar to our study. Leung (2011) applied 4, 5, 6, and 11-point Likert scales in their study and did not find a big difference in Cronbach Alpha value and factor loads. In Chang's (1994) and Preston and Colman's (2000) studies, scales with fewer categories in the response set gave higher reliability values. Prior studies have shown that differences in response categories do not change the Cronbach Alpha coefficient much and that scales with fewer response categories offer a higher level of reliability. In our study, when [Figure 3](#) is examined, it is seen that there is not much difference between the reliability levels. However, as the number of categories decreased, reliability decreased, and as the number of

categories increased, reliability increased. In this respect, it can be said that the study results are compatible with the study conducted by Leung (2011).

The closest fit values to the item-model fits determined in the literature were obtained in the 5-point Likert form. The increase in the number of degrees in the Likert response set in the scale forms increased the discrimination. In this study, the item options differentiated and worked better in the 4-point Likert form. The number of items with the least undifferentiated option is in the 4-point Likert form. The 4-point Likert form had the least items with unspecific responses from medical students. Therefore, the 4-point Likert form explained the variance of the scale's measured feature the least, and the item-model fit parameters were not at the level suggested by the literature. However, the 4-point Likert form performed well in terms of working out the item options and obtaining the participants' responses. In the study by Aybek and Toraman (2022), forms of a measurement tool with 3-point, 5-point, and 7-point response sets were tested. The researchers analyzed the data they obtained based on IRT. The results showed no difference between the three forms in terms of "a" parameters, and the 5-point and 7-point response categories were more advantageous regarding test knowledge and reliability functions. However, seven response categories according to OCCs could not be distinguished by the participants. According to the research of Adelson and McCoach (2010) and Aybek and Toraman (2022), scale forms with 5-point response sets work well. Wakita et al. (2012) applied the forms of a scale with 4, 5, and 7-point response sets to 722 students. The researchers analyzed the data based on IRT. The results showed that the number of degrees of the scale affects the psychological distance between the options, especially for the scale with 7 degrees.

In the present study, an examination of the test information functions showed that the 4-point Likert form provides the least information. The 5-point and 6-point Likert forms gave information close to each other. The 5-point and 6-point Likert forms gave higher information than the 4-point Likert forms and lower than the 7-point Likert forms. The most informative form was the 7-point Likert. When the reliability functions were examined, the reliability level of all four forms exceeded 0.80 and were reliable at a level close to each other. The form with the highest reliability was the 7-point Likert form, albeit by a small margin. On the other hand, the 4-point, 5-point, and 6-point Likert forms were similar and had higher internal consistency in a slightly wider theta range compared to the 7-point Likert form. In the study by Aybek and Toraman (2022), test information and reliability functions showed that using the 7-point response category could provide a better advantage over using the 5-point response.

As a result, increasing the number of degrees in the response sets positively affected the level of informing, and the level of variance explained regarding the feature of interest. However, the 4 and 5-point Likert-type forms were also prominent in terms of better discrimination of options, not less advantageous than the 6 and 7-point forms.

5. LIMITATIONS

In the study, all participants were administered the 4-point, 5-point, 6-point, and 7-point Likert forms of the CPMS at different times (leaving the scale items long enough to be forgotten). This way, data of four different forms could have been obtained from 2150 medical school students. However, the vast majority of the participants did not accept participation in all four different forms. This situation prevented some comparisons (such as comparing the scores of each individual in all forms).

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:** Çanakkale Onsekiz Mart University, Scientific Research Ethics Committee, 03.02.2022 dated 03/11 numbered

Contribution of Authors

Murat Tekin: Investigation, Resources, and Writing-original draft. **Çetin Toraman:** Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Ayşen Melek Aytuğ Koşan:** Investigation, Formal Analysis, and Writing-original draft.

Orcid

Murat Tekin  <https://orcid.org/0000-0001-6841-3045>

Çetin Toraman  <https://orcid.org/0000-0001-5319-0731>

Ayşen Melek Aytuğ Koşan  <https://orcid.org/0000-0001-5298-2032>

REFERENCES

- Adelson, J.L., & McCoach, D.B. (2010). Measuring the mathematical attitudes of elementary students: The effects of a 4-point or 5-point Likert-Type scale. *Educational and Psychological Measurement*, 70(5) 796-807. <https://doi.org/10.1177/0013164410366694>
- Aiken, L.R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement*, 43, 397-401.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice-Hall International, Inc.
- Aybek, E.C., & Toraman, C. (2022). How many response categories are sufficient for Likert type scales? An empirical study based on the Item Response Theory. *International Journal of Assessment Tools in Education*, 9(2), 534-547. <https://doi.org/10.21449/ijate.1132931>
- Aytug Kosan, A.M., & Toraman, C. (2020). Development and application of the commitment to profession of medicine scale using classical test theory and item response theory. *Croatian Medical Journal*, 61(5), 391-400. <https://doi.org/10.3325/cmj.2020.61.391>
- Bora, B. (2013). *A study on the applicability of the likert type scales in marketing*. Doctoral Thesis. Sakarya University. Sakarya.
- Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen K., Long J. (Eds.), *Testing structural equation models* (pp. 136-162). SAGE.
- Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29. <https://doi.org/10.18637/jss.v048.i06>
- Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, 23, 323-331.
- Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, 18(3), 205-215. <https://doi.org/10.1177/014662169401800302>
- Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research*, 50(1), 61-104. <https://doi.org/10.1177/147078530805000106>
- DeVellis, R.F. (2003). *Scale development, theory and applications*. SAGE Publications.
- Dunn-Rankin, P., Knezek, G.A., Wallace, S., & Zhang, S. (2004). *Scaling methods*. Lawrence Erlbaum Associates, Inc.
- Flannery, W.P., Reise, S.P., & Widaman, K.F. (1995). An item response theory analysis of the general and academic scales of the self-description questionnaire II. *Research in Personality*, 29(2), 168-188. <https://doi.org/10.1006/jrpe.1995.1010>
- 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>
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2014). *Multivariate data analysis*. Pearson Education Limited.
- Hambleton, R.K. (1994). Guidelines for adapting educational and psychological test: A progress report. *European Journal of Psychological Assessment*, 10(3), 229-244.

- Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, 38, 1217-1218
- Joshi, A., Kale, S., Chandel, S., & Pal, D.K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology (BJAST)*, 7(4), 396-403. <https://doi.org/10.9734/BJAST/2015/14975>
- Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, 6(2), 151-162. <https://doi.org/10.32614/RJ-2014-031>
- Leung, S.O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert Scales. *Journal of Social Service Research*, 37, 412-421. <https://doi.org/10.1080/01488376.2011.580697>
- Likert, R. (1932). A technique for the measurement of attitudes. *Arch Psychology*, 22(140), 55.
- Lord, F.M. (1954). Chapter II: Scaling. *Review of Educational Research*, 24(5), 375-392. <https://doi.org/10.3102/00346543024005375>
- Mariano, L.T., Phillips, A., Estes, K., & Kilburn, R. (2024). *Should survey Likert Scales include neutral response categories? Evidence from a randomized school climate survey*. Working Paper. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/working_papers/WRA3100/WRA3135-2/RAND_WRA3135-2.pdf
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Adv in Health Sci Educ* 15, 625-632. <https://doi.org/10.1007/s10459-010-9222-y>
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. McGraw-Hill, Inc.
- Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104, 1-15. [https://doi.org/10.1016/s0001-6918\(99\)00050-5](https://doi.org/10.1016/s0001-6918(99)00050-5)
- Price, L.R. (2017). *Psychometric methods, theory into practice*. The Guilford Press
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Revelle, W. (2021). *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston, Illinois. R package version 2.1.6. <https://CRAN.R-project.org/package=psych>
- Samejima, F. (2005). *Graded response model in encyclopedia of social measurement*, edit. Kimberly Kempf-Leonard (pp: 145-153). Elsevier. <https://doi.org/10.1016/B0-12-369398-5/00451-5>
- Smits, N., Öğreden, O., Garnier-Villarreal, M., Terwee, C.B., & Chalmers, R.P. (2020). A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement. *Statistical Methods in Medical Research*, 29(4), 1030-1048. <https://doi.org/10.1177/0962280220907625>
- Sodano, S.M., Tracey, T.J.G., & Hafkenscheid, A. (2014) A brief Dutch language impact message inventory-circumplex (IMI-C Short) using non-parametric item response theory. *Psychotherapy Research*, 24(5), 616-628. <https://doi.org/10.1080/10503307.2013.847984>
- Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680
- Thomas, H. (1982). IQ interval scales, and normal distributions. *Psychological Bulletin*, 91, 198-202
- Torgerson, W.S. (1958). *Theory and methods of scaling*. John Willey & Sons, Inc.
- Warner, R.M. (2013). *Applied statistics, from bivariate through multivariate techniques*. SAGE Publications, Inc.
- Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement*, 72(4), 533–546. <https://doi.org/10.1177/0013164411431162>

- Wong, C.-S., Chuen, K.-C., & Fung, M.-Y. (1993). Differences between odd and even number of response scales: Some empirical evidence. *Chinese Journal of Psychology*, 35, 75-86.
- Wu, H., & Leung, S.O. (2017). Can Likert Scales be treated as interval scales? A simulation study. *Journal of Social Service Research*, 43(4), 527-532. <https://doi.org/10.1080/01488376.2017.1329775>
- Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3),187-213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>
- Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

Educating non-specialized audiences about seismic design principles using videos and physical models

Mauricio Morales-Beltran^{1*}, Ecenur Kızılörenli¹, Ceren Duyal²

¹Yaşar University, Faculty of Architecture, Department of Architecture, İzmir, Türkiye

²Eindhoven University of Technology, Department of Built Environment, Eindhoven, The Netherlands

ARTICLE HISTORY

Received: Feb. 22, 2024

Accepted: June 28, 2024

Keywords:

Educational media,
Earthquake awareness,
Qualitative assessment,
Knowledge survey,
Architectural education.

Abstract: The prevalence of self-construction practices in Türkiye has resulted in a building stock whose earthquake resilience is highly uncertain. To mitigate the potentially devastating impact of anticipated large earthquakes, one viable approach is to increase earthquake awareness among builders themselves. However, these builders lack formal engineering training and are ordinary citizens. Therefore, the challenge lies in devising visual teaching methods, such as short videos, to explain complex seismic phenomena in a comprehensible manner. This paper introduces the use of educational media tailored for non-specialized audiences, encompassing regular citizens and students without engineering backgrounds. These videos are based on experiments conducted with physical models on a homemade shake table. They focus on key factors influencing the seismic response of multi-storey buildings and highlight common design and construction errors that lead to building damage. To assess the effectiveness of this approach, we conducted a workshop with junior architecture students, followed by post-workshop qualitative assessments through knowledge surveys and interviews. The findings indicate that while single-topic videos were effective learning tools for students without prior knowledge of seismic building design, students found models particularly useful for explaining specific concepts such as torsional behavior, the role of diaphragms, and the performance of non-structural components. However, despite positive feedback on the effectiveness of model testing, students generally did not perceive significant knowledge acquisition in model construction. Ultimately, the accessibility of freely available videos, coupled with their enhanced educational value, makes them effective tools for raising seismic awareness in communities vulnerable to future earthquakes.

1. INTRODUCTION

Due to widespread self-construction practices in Türkiye over recent decades, the actual earthquake resistance of existing buildings is uncertain (Dener, 1994; Green, 2008; Iban, 2020). Unfortunately, these practices often result in a lack of compliance with building codes, which has been a major contributor to widespread building damage in both past and recent earthquakes in Türkiye (Binici et al., 2022; Hussain et al., 2023; Yakut et al., 2022). Considering that a

*CONTACT: Mauricio Morales-Beltran ✉ mauricio.beltran@yasar.edu.tr 📍 Yaşar University, Department of Architecture, Üniversite Caddesi No:37-39 Bornova 35100, İzmir, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

significant portion of the building stock does not comply with codes, extensive damage is expected when moderate-to-large earthquakes strike urban areas in the near future.

The current high vulnerability of Türkiye's building stock has prompted calls for immediate action to mitigate seismic risk, such as assessing existing buildings. However, due to the large number of housing units in Türkiye, this cannot be achieved immediately but only with a small number of buildings (Binici et al., 2022). Long-term measures to prevent substandard construction practices in the future include implementing a system to prosecute negligent builders and establishing compulsory registration for builders and contractors. Yet, given the long-term nature of the self-construction problem and its complex interplay with social and economic factors, a more effective solution may be focusing efforts on empowering self-builders to actively reduce cities' vulnerability to natural disasters (Green, 2008). Achieving this requires educating citizens, both current and future, on the key aspects of building seismic behavior.

1.1. Educating Non-Specialized Audiences in Earthquake-Resistant Design of Buildings

Education in seismic-resistant construction for self-builders would prioritize individual action in hazard prevention. This entails simplifying current seismic codes, which can be complex for non-specialists but have been shown to significantly reduce earthquake-related disasters (Spence, 2004). By expanding residents' knowledge of seismic design and construction techniques, they will not only be aware of their homes' vulnerability but also understand how to mitigate it (Green, 2008). Importantly, such education can demonstrate to self-builders that using proper reinforcement detailing and concrete ratios can enhance a structure's seismic resistance without significantly increasing material costs (Green, 2008). Furthermore, education can deter the proliferation of *quick and cheap* construction approaches, where uneducated constructors prioritize selling apartments quickly with ostentatious but inexpensive materials (Dener, 1994).

This paper suggests using educational videos to impart knowledge about building seismic behavior to non-specialized audiences. These videos aim to bridge the gap between professional engineering knowledge and regular citizens, some of whom may consider building their own homes. Originally intended for self-builders, including models in the videos introduces a pedagogical dimension suitable for educational environments like schools and universities. Thus, the term "non-specialized audience" expands to encompass all students. This expansion aligns with calls for earthquake education programs to be integrated into school and community curricula, thereby enhancing ordinary citizens' disaster awareness (Simonacci & Gallo, 2017).

The purpose of this paper is to evaluate the effectiveness of these educational media as both a learning tool and in raising seismic awareness, through qualitative assessments conducted with university students. The selected cohort, a second-year architecture class with no prior seismic design knowledge, provides a controlled environment to evaluate the validity of the learning method and the strategies that enhance its effectiveness.

2. METHODS

2.1. Development of Educational Videos for Seismic-Resistance Design Education

Utilizing videos as a tool for learning enhancement has been a longstanding practice due to their ability to captivate students' attention (Bravo et al., 2011). Low-cost instructional videos, defined as brief promotional videos with specific educational goals, are widely used to increase student engagement and motivation. These videos can be swiftly created, combined, or embedded into course materials with minimal resources, addressing common challenges such as budget constraints and time limitations. The process of incorporating videos into teaching materials is streamlined, facilitating efficient integration into educational settings.

2.1.1. Educational videos using scale-down models

Transforming educational videos into effective learning tools requires critically developing the associated video content from an educational perspective. Given the aim of explaining how structures respond to seismic accelerations, these videos rely on the use of scaled-down models. These models are widely employed in architectural education, with well-known advantages for teaching structural and seismic design principles (Ji & Bell, 2000; Morales-Beltran & Yıldız, 2020).

According to Ji & Bell (2000), physical models make structural concepts and principles more observable and tangible, leading to better student understanding and attention. This approach not only captures students' attention but also promotes deeper understanding. Additionally, manipulating models encourages students to construct their own meaning in acquiring knowledge, rather than memorizing information from other sources such as lecture notes (López et al., 2022). Through this technique, students engage in both surface and deep learning. Research indicates that students constructing their meaning can simultaneously lead to surface learning (e.g., memorizing model behavior) and deep learning (e.g., connecting the model's behavior to principles learned in lessons) (Biggs & Tang, 2011). Surface learning involves memorization, while deep learning focuses on understanding concepts, their reasoning, and their connections with prior knowledge. Deep learning is essential for architecture students to express their ideas and knowledge in their designs (Gunasagaran et al., 2021). Based on this information, we anticipate that through the process of model-making and reviewing the knowledge they initially acquired, students will understand, reinforce, and retain what they have learned in the long term.

The models in the videos are simplified representations of the most common residential building typology in Türkiye: multi-storey reinforced concrete (RC) frame buildings with infill walls (Gulkan et al., 2002). While the seismic behavior of such buildings is complex, simplifying the concepts into cause-and-effect relationships aids in conveying a general understanding to lay audiences. By simulating behaviors based on simple inputs (e.g., force), the models offer a visual representation of seismic phenomena without overwhelming complexity. Moreover, these models provide insights into potential building behaviors during earthquakes without directly referencing existing structures, alleviating concerns among viewers regarding the state of their own buildings. Finally, to ensure ease of replication for educational purposes, instructions for fabricating the models are provided as part of the learning experience.

Using physical models for the videos involves the process of fabricating these models by the audience, fostering active learning. While a single model (or set of related models) can be fabricated and tested during a single lesson or workshop, online how-to videos can complement this work, providing the benefits of blended learning. Also known as hybrid or mixed-mode learning, blended learning is the integration of face-to-face and online learning to enhance the classroom experience and extend learning through the innovative use of information and communications technology (Blackmore et al., 2010; Bregger, 2017; Iskander, 2007; Napakan et al., 2009). Blended strategies enhance student engagement and learning through online activities, reducing lecture time (Watson, 2008). Additional advantages include increased student retention, flexibility to study at a convenient time and place (Partridge et al., 2011), and an improved overall learning experience and outcome (Hajhashemi et al., 2016).

2.1.2. Design and fabrication of physical models

The video topics cover various parameters influencing the seismic performance of multi-storey RC residential buildings with infill walls, including ground motions, seismic-resistant configurations, and non-structural elements (see [Table 1](#)). The number of videos corresponds

to the number of topics, plus two addressing fabrication issues. Topics were selected based on essential content recommended for earthquake-resistant building courses (Charleson, 2018). Each video features several models designed and fabricated to demonstrate key concepts, with model testing serving as a central component.

Table 1. Topics addressed in the videos and corresponding models.

#	Topic	Issues	Target	Models
1	Buildings' Natural Period	Ground motions & buildings Buildings' natural period (T)	Effect of seismic waves on buildings Buildings with different heights Buildings with different masses	None 1-, 8-, and 16-storey* Two 8-storey*
2	Lateral Resistant Systems	Moment frames Shear Walls Braced Frames	Effects of the connections, bracings and walls	2-storey*
3	Diaphragms and Openings	Role of the diaphragms Suitable openings placement	Flexible slab w/o penetrations	Two 2-storey forming a 3-bay structure
4	Building Configuration Irregularities - Part 1: Torsion	Torsion	Torsion due to eccentricity - Centre of Resistance Torsion due to eccentricity – Centre of Mass	8-storey*
5	Building Configuration Irregularities - Part 2: Irregular Plans & Pounding	Re-entrant corners Pounding	L-shape plans & seismic gaps Pounding & seismic gaps	1-storey* and 3-storey 3-bay
6	Building Configuration Irregularities - Part 3: Soft Storey & Short Columns	Soft stories Short columns	Soft stories / ground + infill walls Short columns / deep foundation hole & rising foundation	2-storey* 8-storey*
7	Non-structural Elements	Infill & partition walls	Role of infills - non-structural damage	2-storey*
8	Fabrication of the Shake Table	Do-it-yourself	Materials, construction & assembly process	None
9	Fabrication of the Building Models and variations	Do-it-yourself	Materials, 3D printed pieces, construction & assembly process; loading	All

* single-bay structures

The underlying assumption when using physical models to facilitate the understanding of the dynamics behind the seismic performance of buildings is that the model behaves as a full-scale building. Therefore, models are designed at a 1/60 scale, representing two actual floors per storey. While most videos feature a generic 8-storey model (Figure 1), additional models with varying storeys were utilized to highlight specific issues. Components for fabrication include 4mm wooden sticks for columns, 3mm cardboard for semi-rigid diaphragms, and customized 3D-printed pieces for connections (Figure 2). All connections are designed for easy assembly and disassembly, eliminating the need for adhesives. Comprehensive information on these models can be found in Morales-Beltran et al. (2021).

Figure 1. Testing of generic 8-storey models (each level representing two actual building floors): before moving the shake table back and forth (left) and freeze-frame during the testing, displaying models differentiated lateral deformations (right).

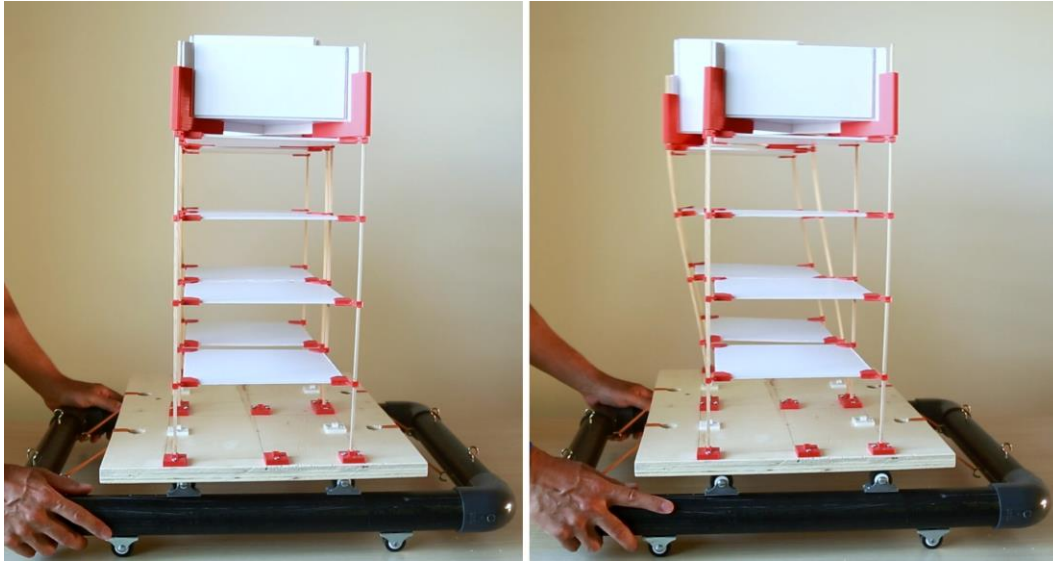
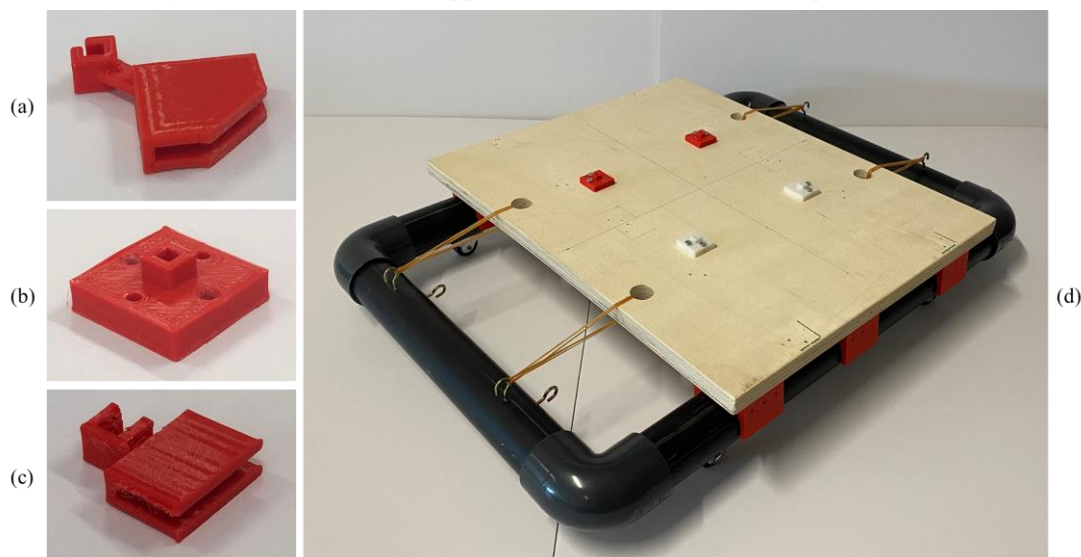


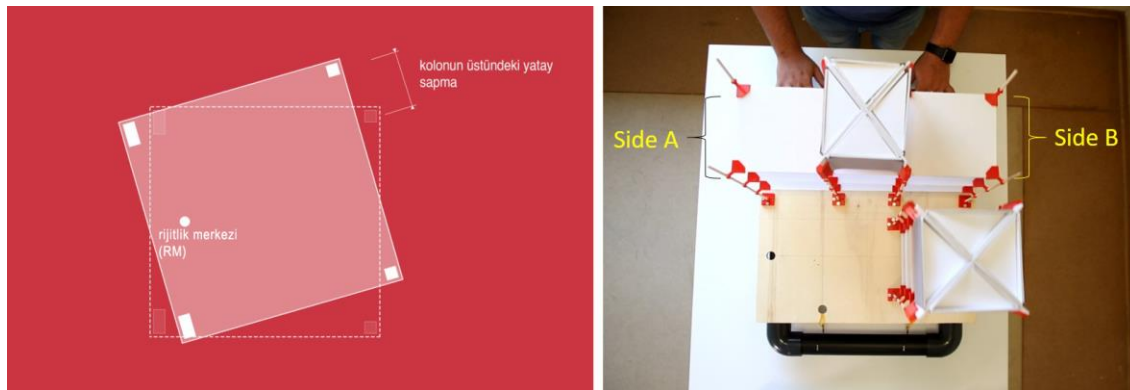
Figure 2. image of main model components: (a) corner connections to attach a 3mm cardboard to a 4mm stick; (b) foot connections; (c) clippers, and (d) shake table components.



2.1.3. Video production

Video shoots were conducted in standard classrooms, requiring minimal equipment. Each video focused on a few key aspects to facilitate student comprehension. This approach facilitates student comprehension by presenting information in manageable pieces, allowing them to control the flow of information (Brame, 2016). Once the recordings were completed, additional images and animations were incorporated to enhance the visual narration (Figure 3). Following the principles of signaling, keywords, texts, color changes, and symbols were added in various ways to highlight important information and direct viewers' attention (İbrahim et al., 2012). Additionally, videos were kept as short as possible to maintain students' interest. Research suggests that videos shorter than 6 minutes are more successful in capturing viewers' attention, while longer videos tend to lose it.

Figure 3. Examples of visual enhancements: a GIF in high contrast used to reinforce conceptual understanding (left) and additional legends over the freeze-framed video to increase clarity of the explanation (right).



2.1.4. Do-it-yourself (DIY) videos

Two supplementary videos were created to guide individuals in replicating the models and tests showcased in the educational videos. These DIY videos aim to empower instructors, architects, engineers, and builders to utilize the models for educational purposes in various settings. Digital files necessary for 3D printing components are freely accessible online, facilitating easy replication of the models.

2.2. Qualitative Analyses Using Surveys and Interviews

After observing that the videos were being watched, our goal was to understand what viewers actually learned during and after watching them. Most importantly, we aimed to determine if they acquired the expected knowledge and to what extent. To answer these questions, we conducted qualitative assessments of the video-based learning experiences, including a workshop, surveys, and interviews with architecture students. Conducting the workshop with university students offers control, monitoring, and consistency across different surveys compared to a public audience. Moreover, working with architecture students ensures they gain a deep understanding of seismic-resistant design. This knowledge is crucial because:

- Architectural decisions, especially early ones concerning building shape and configuration, significantly influence a building's seismic performance. Architects who design with an understanding of these effects can prevent irregularities and discontinuities, thereby avoiding extended damage or collapse (Charleson, 2018; Morales-Beltran & Yildiz, 2020; Özmen & Ünay, 2007).
- Architects often have direct involvement in the construction process, including acting as contractors (Dener, 1994).

2.2.1. Workshop

Students typically have a basic understanding of structural design by the end of their architectural education, but they may lack knowledge of seismic design principles. Therefore, the workshop was aimed at second-year architectural students at Yasar University in Izmir, Türkiye. These students were part of a course focusing on basic structural principles and had not been exposed to seismic design concepts due to the previous year's online education format caused by COVID-19. The workshop spanned three weeks and was integrated into the weekly 3-hour practice sessions of the course. Sixty-five students organized themselves into teams of 4-6 members. Teams received weekly assignments without prior knowledge of the specifics of the exercise (Table 2).

Table 2. Main activities and assessments developed during the workshop.

Week	Workshop	Materials	Assessment
1	Students took the baseline survey (S1) and then watched the seven* seismic-related videos in the classroom	None	Baseline S1 – Knowledge survey
2	Students took the survey (S2-1) and then began studying the assigned video and models. By the end of the session, they took another survey (S2-2)	3D connections were distributed, while students were expected to bring their own materials (cardboard, wooden sticks, etc.)	S2-1 – Knowledge survey; S2-2 – Knowledge and Validation surveys
3	Presentation and testing of the models. After that, students took the last survey (S3)	Models and beans (or similar) acting as masses	S3 – Knowledge and Validation surveys

* The other two videos describing fabrication issues were excluded.

During the second week of the workshop, teams were provided with 3D printed components to assemble their models. The number of components provided was insufficient to complete the model, testing whether students would 3D print the missing pieces or find alternative solutions. Monitoring their responses provided insights into how other students might handle similar challenges. The workshop was experimental, and participation was not graded based on performance, but merely on attendance.

2.2.2. Knowledge and validation surveys

The survey aimed to assess the seismic-related knowledge gained by students after:

- Watching seven educational videos (about 6-9 minutes each)
- Building and testing only one of the models utilized in the videos

A longitudinal panel survey was prepared to evaluate students' knowledge throughout the workshop. Eight months after the workshop, interviews with selected students were conducted based mainly on open-ended questions.

Knowledge surveys were employed to assess the learning process. These surveys present learning objectives framed as questions that evaluate mastery of specific content areas (Nuhfer & Knipp, 2003). Rather than providing direct answers, students indicate their perceived ability to answer using predefined scaled options (Wirth & Perkins, 2005). Our survey utilized a three-point scale (see Table 3), and consisted of 40 questions focusing on knowledge retention and comprehension questions structured around Bloom's cognitive domains (See Appendix 1 for full questionnaire).

Table 3. Responses available to students for answering questions on the knowledge survey. Source: (Wirth & Perkins, 2005).

#	Answer
1	I do not understand the question, I am not familiar with the terminology, or I am not confident that I can answer the question well enough for grading purposes at this time
2	I understand the question and a) I am confident that I could answer at least 50% of it correctly, or b) I know precisely where to find the necessary information and could provide an answer for grading in less than 20 minutes
3	I am confident that I can answer the question sufficiently well-enough for grading at this time

Despite the widespread acceptance of knowledge surveys as effective learning assessment tools, there is a recognized concern about their reliability as indicators of student understanding

(Wirth & Perkins, 2005). In their study, Wirth and Perkins compared knowledge surveys with students' exam scores and final grades to assess the reliability of the primary survey. To ensure the validity and consistency of our survey results and mitigate potential biases—such as students feeling compelled to demonstrate confidence—we included validation surveys in the second part of the workshop (see Table 2).

In these validation surveys, students were instructed to choose one question per section (each section corresponding to a specific video and containing up to 6 questions) and answer it as they would in a regular test. Therefore, each validation survey contained only seven questions. Subsequently, instructors evaluated and categorized these responses using the same three-point scale as the knowledge survey (Table 3), establishing *key answers*. The numerical difference between the key answers and students' responses (Δ) indicates the level of agreement or discrepancy between the surveys. Additionally, the Δ value serves as a measure of the reliability and credibility of students' answers.

2.2.3. Interviews with students

The purpose of conducting interviews was to gather comprehensive data on the effectiveness of visual media and models as learning tools for understanding seismic design principles. Semi-structured, in-depth interviews were chosen as the method of data collection due to their flexibility in adapting to a predefined set of open-ended questions and allowing for spontaneous follow-up questions during interactions between interviewers and interviewees (DiCiccio-Bloom & Crabtree, 2006). In-person interviews were preferred because they enable interviewers to capture participants' verbal and non-verbal cues, which often provide insights that can lead to further exploration (Adeoye-Olatunde & Olenik, 2021).

The overarching aim of using individual semi-structured interviews was to provide a clear and focused structure for discussions while also allowing space for participants to express their individual perspectives. This approach facilitated gathering diverse data on similar topics from different participants (Kallio et al., 2016). The interviews were based on semi-open questions organized into four main sections (full description in Appendix 3: Interview Questions):

- Video: Recollection of what participants remembered from the videos and what aspects helped them understand.
- Working: Assessment of how influential the videos were in the process of constructing the models.
- Testing: Comparison of participants' testing processes with those demonstrated in the videos.
- Learning: Reflection on what participants perceived they had learned and areas where their understanding might still be lacking.

Additionally, the interview included three supplementary parts: soliciting suggestions for improvement to encourage forward thinking, exploring participants' learning processes to foster reflection, and a brief survey. In this survey, participants were asked to consider themselves as active learners tasked with teaching other students the content covered in the knowledge surveys. They indicated whether they would use a model or a video for each question to facilitate teaching.

The in-person interviews were conducted between January and February 2023, more than eight months after the surveys were administered. Fourteen students, two from each video group, were selected based on their high scores in the surveys, specifically those showing the best alignment between their validation survey responses and the key answers. Eleven students (St01-St11) accepted the invitation. They were informed in advance that the interview would focus on their previous coursework in statics, without specific reference to the workshop involving videos and models. This approach aimed to prevent students from preparing by

revisiting the videos. Interviews were conducted in both English and Turkish based on students' language preference and comfort level in expressing their thoughts.

3. RESULTS

3.1. Performance of The Videos in The Youtube Channel

All nine videos were uploaded to a YouTube channel named “Earthquakes & Buildings” (BAP103-Deprem & Binalar, 2021) between June 2021 and February 2022. A year later, the channel had gained 221 subscribers and accumulated approximately 3,600 views of the videos. By May 2023, the channel's subscriber count had increased to 858, with the videos collectively receiving about 16,300 views. Excluding the do-it-yourself videos, a significant percentage—ranging between 74% and 90%—of these 16,000+ views occurred only after the Kahramanmaraş Earthquakes of February 6, 2023 (Table 4). This increase in viewership can be interpreted as people seeking answers, particularly amid uncertainties regarding building collapses and construction quality in the aftermath of the earthquakes. The fact that the most-watched videos maintained an average viewing time of over 50% indicates that the videos successfully held the viewers' attention.

Table 4. List of videos of the “Earthquakes & Buildings” YouTube channel by May 2023, organized by number of total views.

#	Video	Duration	Average Watching Time	Uploaded	Views	
					Total*	After 06/02/2023
6	Building Configuration Irregularities - Part 3: Short Columns & Soft Storey	07:19	51%	01/2022	4780	90%
1	Natural Period of Buildings	05:27	54%	06/2021	3644	83%
4	Building Configuration Irregularities - Part 1: Torsion	07:48	51%	10/2021	1866	74%
7	Non-structural Elements: Infill Walls	05:17	51%	02/2022	1339	77%
2	Lateral Force-Resistant Systems	09:14	47%	10/2021	1269	79%
5	Building Configuration Irregularities - Part 2: Irregular Plans & Pounding	09:20	49%	12/2021	1189	74%
3	Diaphragms and Openings	08:27	47%	02/2022	905	77%
8	Fabrication of the Shake Table	07:00	22%	06/2021	867	36%
9	Fabrication of the Models	07:09	23%	01/2022	402	22%

*As of 24/05/2023

3.2. Survey Results

Sixty-five students initially participated in the workshop, but active involvement and consistent survey responses were maintained by 51 students. Hence, the results are based on these 51 surveys, focusing on knowledge increase and survey consistency.

Since the baseline (S1) and S2-1 are knowledge surveys, the results consider the answers given by all 51 students to all questions. After the teams were assigned specific videos, they separately focused on each video-related work. Since S2-2 and S3 are both knowledge and validation surveys, i.e. students choose a question to be answered, their results separately account for the answers given only by the teams working on each of the seven videos. Consequently, the scores per video in S1 and S2-1 were computed using 51 answers, whereas the scores per video in S2-2 and S3 were computed using only between eight and ten answers.

3.2.1. Knowledge increase

3.2.1.1. Most frequently selected questions for answering. The validation survey required students to select and answer one question from each of the seven topics covered in the knowledge survey. Table 5 highlights the questions most frequently chosen for answering in each section.

In the baseline survey (S1), questions such as “What is an earthquake?”, “What is center of mass”, and “What is the most common type of non-structural infill wall used in Türkiye” (questions 1, 19, and 36 respectively) received relatively higher scores, suggesting existing prior or common knowledge among the students. The consistent average scores across subsequent surveys support this observation. For the other four videos, there was a notable increase in scores in S2-1 (taken after watching all videos). It indicates that the videos contributed significantly to the students' understanding.

The average scores in S2-2 and S3 reflect answers only from students who constructed the model(s) related to the assigned video. Each video was assigned to two teams with four or five members each, resulting in between eight and ten students per video topic by the final survey.

While the score variations in Table 5 do not follow a distinct pattern, the significant variation observed in question 17's scores is noteworthy. The increase in S2-1 compared to the baseline survey indicates that the videos helped most students understand the optimal location for openings in a diaphragm. However, the lower score in S2-2 suggests that fewer students assigned to study that specific video had a strong grasp of the topic. The subsequent increase in S3 compared to S2-2 demonstrates that constructing models enhanced understanding among the students working on that video.

Table 5. Average score of the most frequently selected to-be-answered question in validation surveys.

Video		Knowledge Survey		Average score			
#	Topic	No.	Question	Baseline*	S2-1*	S2-2**	S3**
1	Natural Period of Buildings	1	What is an earthquake?	2.75	2.71	2.60	2.60
2	Lateral Force-Resistant Systems	7	Why vertical continuity is fundamental to provide buildings with adequate resistance to earthquakes?	1.75	2.08	2.30	2.30
3	Diaphragms and Openings	17	From a seismic-resistance perspective, where is the best location to make openings in the diaphragm?	1.18	2.24	1.33	2.50
4	Torsional Behaviour	19	What is Centre of Mass?	2.49	2.65	3.00	2.88
5	Irregular Plans & Pounding	26	Why irregular plan layouts can be potentially dangerous during earthquakes?	1.90	2.41	2.30	2.60
6	Short Columns & Soft Storey	29	What is a short column?	1.86	2.45	2.57	2.71
7	Non-structural Infill Walls	36	What is the most common type of non-structural infill wall used in Türkiye?	2.06	2.26	2.50	2.17

Score ranges between 1 and 3

* All students' answers in the knowledge surveys

** Considering only the answers from students who worked with the specific video topic

3.2.1.2. Least frequently selected questions for answering. Examining the least frequently chosen questions provides insights into the subjects that challenged students' understanding the most. The variations across surveys offer clear indications of the students' learning progression. As depicted in Table 6, positive variations between S2-1 and the baseline survey for all videos suggest that students made learning gains after watching the videos. Similarly, between S3 and S2-2, positive variations are observed in 6 out of the 7 videos, indicating enhanced learning after constructing and testing models.

Notably, the substantial increase in S3 compared to S2-2 for question 23 suggests that constructing models related to video #4 significantly improved students' understanding of buildings' torsional behavior. The exception is question 34, concerning soft storey mechanisms, where the negative variation between S3 and S2-2 may indicate that constructing models did not effectively enhance students' comprehension of this topic.

Table 6. Average score of the least frequently selected to-be-answered question in validation surveys.

Video		Knowledge Survey		Average score			
#	Topic	No.	Question	Baseline*	S2-1*	S2-2**	S3**
1	Natural Period of Buildings	3	What are ground motions?	2.12	2.37	2.40	2.40
2	Lateral Force-Resistant Systems	12	What are the most common configurations of braced frames?	1.20	1.98	2.00	2.40
3	Diaphragms and Openings	14	What is the role of a diaphragm in providing seismic resistance to buildings?	1.22	1.96	1.67	2.00
4	Torsional Behaviour	23	How does eccentricity affect the torsional behaviour of a building during an earthquake?	1.18	1.90	1.63	2.75
5	Irregular Plans & Pounding	28	In practice, how wide should the seismic gap be?	1.37	2.10	2.40	2.60
6	Short Columns & Soft Storey	34	Why soft storey mechanisms are dangerous?	1.37	2.10	2.43	2.29
7	Non-structural Infill Walls	38	What type of damage appears when infill walls resist in-plane inertia forces?	1.25	1.94	1.83	2.33

* Considering all answers.

** Considering only answers from students who worked with the specific video topic.

3.2.1.3. Average score variation (Δ) between surveys. The average score variation (Δ) between surveys provides a measure of the changes in students' understanding over time. Positive variations typically indicate learning gains. Survey S2-1, conducted after students watched the videos, shows the largest positive variations compared to the baseline survey (Table 7), indicating learning through video instruction. The greatest improvements between S2-1 and S1 relate to questions associated with video #3, followed by those of video #4 and #2.

Table 7. The seven highest increase in the average difference of scores (Δ) between S2-1 and Baseline.

Video		Knowledge Survey		Δ S2-1 – Baseline
#	Topic	No.	Question	
3	Diaphragms and Openings	17	From a seismic-resistance perspective, where is the best location to make openings in the diaphragm?	1.06
		16	From a seismic-resistance perspective, where is the worst location to make openings in the diaphragm?	1.02
4	Torsional Behaviour	21	What is Centre of Resistance?	0.96
2	Lateral Force-Resistant Systems	8	Why seismic resistance must be provided in both orthogonal plan directions?	0.90
4	Torsional Behaviour	20	What is Stiffness?	0.88
6	Short Columns & Soft Storey	31	Which design solutions help to prevent short columns mechanisms in buildings during earthquakes?	0.86
3	Diaphragms and Openings	13	Which structural elements in buildings are considered as diaphragms?	0.84

Survey S3, conducted after students constructed, tested, and presented their models, reflects learning from hands-on modeling, testing, and discussions with instructors. The largest improvements between S3 and S2-2 (Table 8) are observed for videos #3 and #4, suggesting that students' understanding of these topics significantly improved after engaging with physical models.

These findings highlight the effectiveness of both video-based instruction and hands-on modeling in enhancing students' comprehension of seismic design principles. The positive score variations across surveys underscore the benefits of combining theoretical instruction with practical application in educational settings.

Table 8. The seven highest increase in the average difference of scores (Δ) between S3 and S2-2.

Video		Knowledge Survey		Δ S3 – S2-2
#	Topic	No.	Question	
4	Torsional Behaviour	22	What is Eccentricity?	1.25
3	Diaphragms and Openings	17	From a seismic-resistance perspective, where is the best location to make openings in the diaphragm?	1.17
4	Torsional Behaviour	23	How does eccentricity affect the torsional behaviour of a building during an earthquake?	1.13
		20	What is Stiffness?	0.63
		21	What is Centre of Resistance?	0.63
7	Non-structural Infill Walls	38	What type of damage appears when infill walls resist in-plane inertia forces?	0.50
3	Diaphragms and Openings	15	Why openings might jeopardize the structural integrity of a diaphragm?	0.50

3.2.1.4. Average score variation (Δ) between key answers. To calculate the factored average score (S_f) for each question j in surveys S2-2 and S3, we use Equation 1:

$$S_{f_j} = (\sum_{i=1}^n s_{i,j})/n \quad (\text{equation 1})$$

Where:

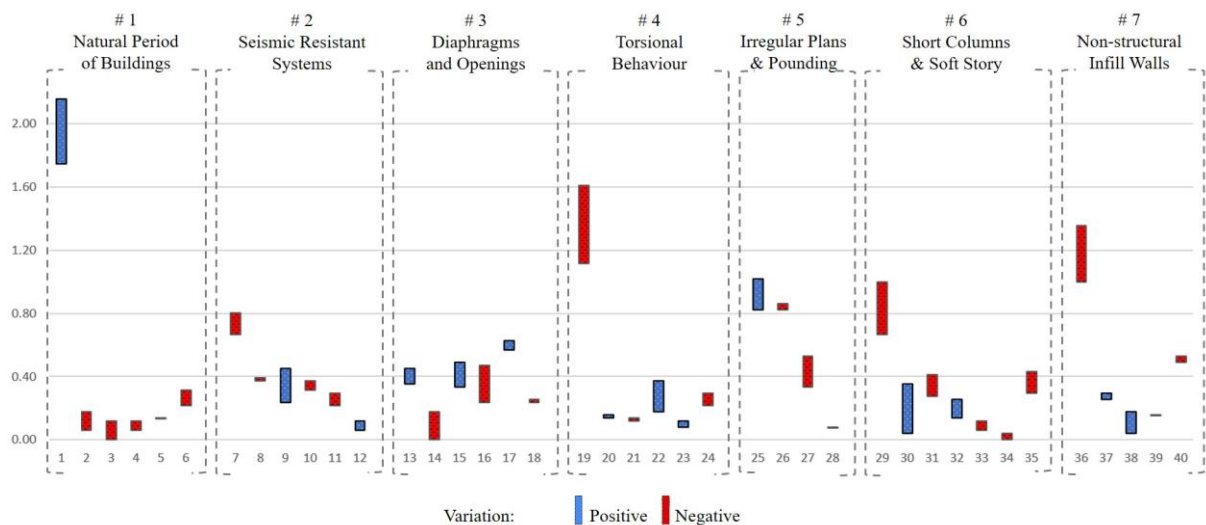
- S_f is the factored average score for question j ,
- s is the score of each answer i given to question j ,
- n is the total number of answers given (in this case, 51).

This computation allows us to assess positive or negative variations between the scores of S3 and S2-2. Positive variations indicate increased learning, while negative variations suggest the opposite.

In Figure 4, which displays these factored scores per survey, questions such as “what is an earthquake” (1), “what is the centre of mass” (19), and “what is the most common type of non-structural infill wall used in Türkiye” (36) received the highest number of responses. However, questions 19 and 36 show negative variations, indicating that the average scores in S3 were lower than in S2-2 for these questions.

This analysis helps in understanding how students' understanding evolved between the validation surveys S2-2 and S3, particularly for questions that received significant responses and exhibited notable score variations.

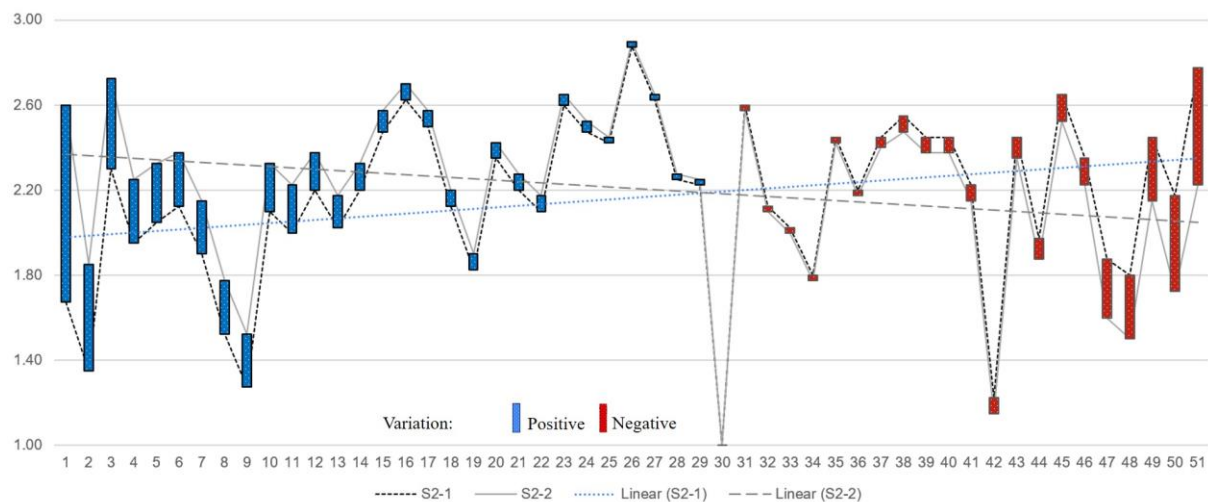
Figure 4. Variations of the average test responses (key answers) between S2-2 and S3 surveys. Higher vertical placement indicates higher number of answers, while length of bars indicates the degree of variation.



3.3.1. Surveys' consistency assessment

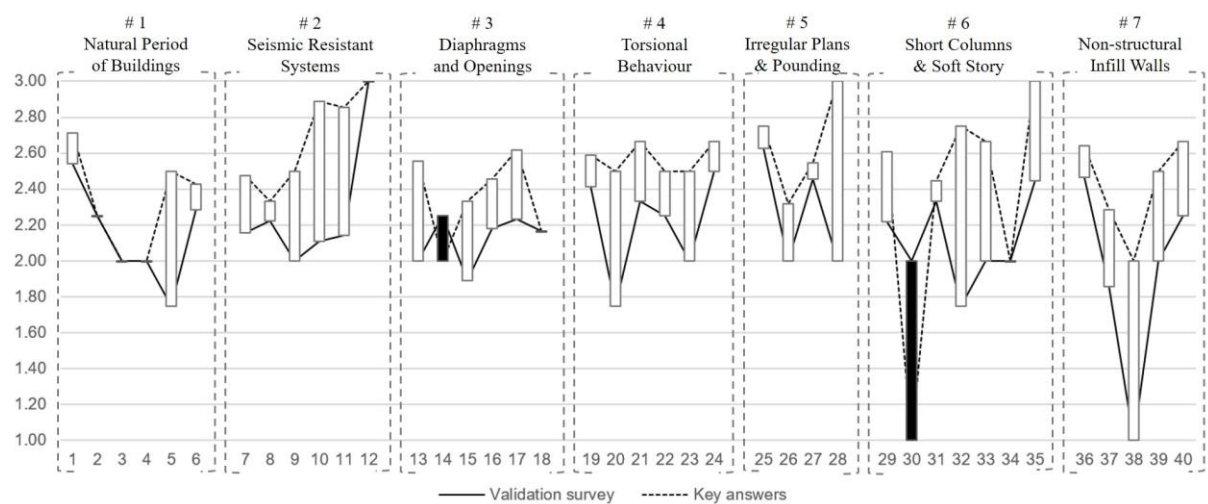
3.3.1.1. Score variations between S2-2 and S2-1. S2-2 was administered shortly after S2-1, with the key difference being that students were aware their perceived knowledge levels would be tested during S2-2. Therefore, variations between their responses in these surveys serve as indicators of students' credibility and, consequently, the reliability of the assessment. In Figure 5, average scores per student from both surveys are arranged in descending order, with the highest positive and negative scores at the right and left ends, respectively. The slight increase in S2-2, with an average score of 2.21 compared to 2.16 in S2-1, reflects the expected outcome that students would perform slightly better after engaging with the videos. The overall average variation of 0.04 shows that 29 answers scored higher in S2-2 ($\bar{x} = 0.18$), while 21 scored lower ($\bar{x} = -0.14$).

Figure 5. Average score per student in S2-1 and S2-2 organizing in decreasing order.

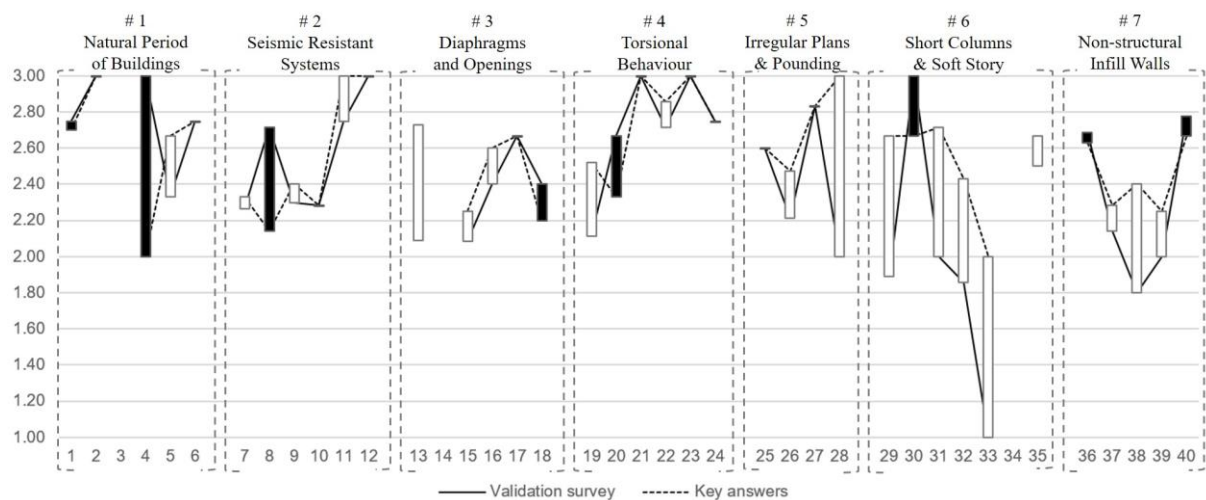


3.3.1.2. Score variations between validation surveys and key answers. Analyzing the average scores of both key answers and students' responses by each video reveals significant discrepancies between students' perceived knowledge (validation survey) and their actual understanding (key answers). Trends observed in S2-2 (Figure 6) indicate that in most questions, students underestimated their knowledge. A notable example is question 38 (“What type of damage appears when infill walls resist in-plane inertia forces?”), where students’ average score was 2.0, while they rated their own understanding at 1.0—the lowest possible score. Conversely, in question 30 (“In which situations do short column mechanisms tend to appear in buildings?”), students displayed overconfidence in their knowledge. There were four matches (equal key and students’ scores) in S2-2, which doubled in S3.

Figure 6. Average score variations between the validation survey and key answers after S2-2.



Moreover, students' overconfidence is apparent in the trends observed in S3 (Figure 7). For instance, in question 30, both the average scores from validation surveys and key answers increased compared to S2-2, indicating students still perceived themselves to know more than they actually did. However, question 4 (“How do ground motions affect buildings?”) showed a slight decrease in scores, and question 33 (“What is a soft storey?”) saw a decrease from 2.0 to 1.0 in validation surveys (students’ perceived knowledge) and from 2.67 to 2.0 in key answers (actual knowledge).

Figure 7. Average score variations between the validation survey and key answers after S3.

3.4. Interview Results

The main insights gathered from the interviews are presented below, categorized based on the students' responses to four main aspects: video, working, testing, and learning.

3.4.1. Techniques to improve retention from videos

When discussing aspects they liked about the videos, students primarily highlighted:

- **Short duration:** Students appreciated the brevity of the videos, noting that lengthy videos hindered their concentration. They found shorter, clear explanations facilitated better understanding.
- **Single focus:** All students agreed that focusing on one topic per video enhanced clarity.

Regarding attention-grabbing features, students positively evaluated:

- **Use of keywords:** Highlighting keywords in contrasting colors aided in understanding and retention, e.g., “the background was red, and the text was white. It is useful to understand and remember short words. They were helpful.” (St09).
- **Visual differentiation:** Different colors and materials used for different components of the model helped distinguish parts effectively, e.g., “the connections were orange or red, floors were white, and columns were timber. We could differentiate the parts of the model.” (St08).
- **Visual explanations:** Students found visual aids such as drawings, images with explanations, and annotations fundamental in enhancing comprehension, e.g., “sometimes the video stopped, images appeared on the screen, and had explanations on it. I think this was very helpful to keep the focus and to understand.” (St07).
- **Additional use of visuals, such as lines, arrows, and letters:** “Like deformation arrows and displacements: showing torsional effects with arrows, bending deformations of the floors made graphic explanations clear, so I understood better.” (St03).

3.4.2. Retention & learning

All students remembered the main subject of their assigned video. For instance, a student explained, “We were supposed to show the torsion behaviour.” (St01). Some of them could even describe the models involved: “The video was about shear walls. We compared 4-, 8-, and 16-storey building models with and without shear walls.” (St07). However, most students provided vague explanations of the video’s purpose: “With our models, we compared different

possibilities under earthquake effects.” (St07), or “We tested the earthquake resistance of the model in terms of damage.” (St09).

Students appreciated the accessibility of the videos, allowing them to watch repeatedly at any time: “We had access to the videos all the time because they were in YouTube.” (St08). This accessibility led to several benefits:

- Learning technical terms: “I think I have learned the meanings of the terms. I cannot remember now, but at that time, I understood. We watched all the videos three or four times.” (St01).
- Understanding details: “After watching it repeatedly, I discovered details. When I first watched it, I understood the given topic generally. But then when I watched it again, I realized the details.” (St09).
- Increased retention: “At the time we watched the videos in the classroom, we watched them several times. And each time, we answered some questions. After we watched it again and again, I was able to answer those questions easier.” (St04).
- Actual learning: “We have worked on short columns, but we watched and learned all the videos. Of course, we learned things about buildings, stability, and earthquake-resistant buildings. We learned like statics of the buildings or how we can create openings on the slab, like we shouldn’t make an opening in the corner of the slab, etc.” (St08).

3.4.3. Video as source of information for making models

Students found videos instrumental in model construction due to their ability to revisit and review. Additionally, videos contained diagrams and a digital axonometric drawing that “really explained the logic of the model.” (St03). Since there were specific and clear explanations, “making the model was easy.” (St07). Without videos, the assembly part would have been the hardest: “Especially for the connections—we would have gotten confused without the videos.” (St11).

Videos demonstrated model behavior on shake tables, helping students to prepare both the model and testing procedure: “it was easy to make the test when you know how the model is supposed to act under forces in that test” (St07). In addition, videos helped students verify the accuracy of their models: “In the preparation stage, we compared our models with another group. We saw that they were doing some parts differently, so we checked the video again. Then we understood that we had missed those parts so we prepared them again quickly.” (St05).

3.4.4. Main benefits of the whole workshop

All interviewed students expressed a positive evaluation of the workshop, highlighting key learning outcomes:

- Understanding theoretical contents, which sometimes remain unclear to the students: “Sometimes in the statics course, the information seems mostly theoretical for us, but with physical models, it was much clearer than a normal lecture.” (St01).
- Enhanced technical terminology: “While answering the survey questions, I realized that my technical vocabulary has improved with this study.” (St07).
- Integration of technical knowledge: Students noted improved comprehension of structural aspects relevant to architectural design. “For example, in the studio, sometimes instructors commented that L-shaped buildings do not work correctly from a structural point of view. Now I can understand this, and sometimes I can also see that a design will not work too.” (St07).

3.4.5. Suggestions for improving learning with videos

When asked for suggestions to enhance future workshops, students proposed:

- Using case studies: Incorporating real building examples to deepen understanding of complex phenomena like soft storeys or short columns.
- Adding closing summaries: Requesting concise summaries at the end of videos containing technical information.
- Question and answer sessions: Suggesting interactive sessions after video viewings to clarify concepts. This suggestion must be understood in the context of students watching the videos during the class.

3.4.6. Sources to learn from

Students ranked activities of the workshop based on their perceived contribution to understanding video topics:

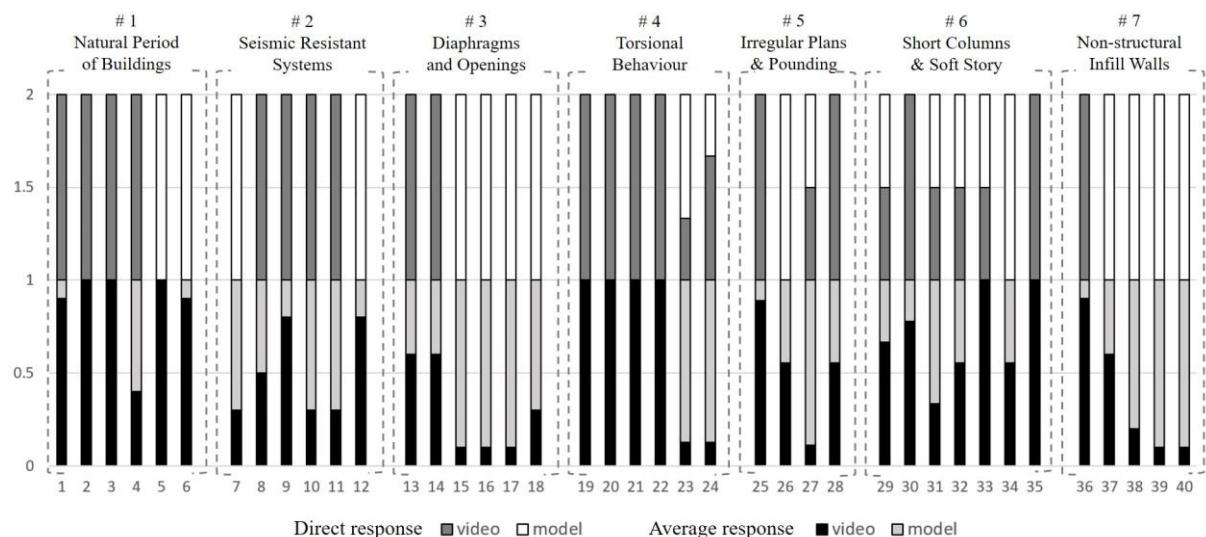
- Watching videos contributed the most, followed by testing models. Five students chose this order, arguing that while testing helped them to get “the main idea” of the model, videos provided more general knowledge.
- Testing models contributed the most, followed by watching videos. The main argument for the second most frequently chosen order was that only through testing was it possible to understand the functional purpose of the model.

Interestingly, *making* models was not widely seen as a significant tool for learning seismic design concepts, possibly due to its perceived commonality in architectural education.

3.4.7. Active role survey

Students were asked how they would use models or videos to teach others about workshop topics. Since each of the interviewed students worked with a particular video, their answers for that particular video’s questions constitute the so-called direct response, separated from the average response—obtained from students who did not work with that video (Figure 8).

Figure 8. Direct (top) and average (bottom) responses to whether students would use videos or models to teach the content of each question.



Results show that:

- Direct and average responses indicated agreement that videos were suitable for explaining most topics except for videos #3 and #7, where models were preferred.

- Direct responses of videos #1 and #2 indicate that at least two questions could be explained using only models, which contradicts the general tendency.

Discrepancies were noted, such as in question 5, where direct responses favored models while average responses did not. Questions 10 and 11 received full agreement for using videos in direct responses but only 30% in average responses. Conversely, questions 6 and 12 received full agreement for using models in direct responses but little in average responses.

4. DISCUSSION

The primary objective of the videos as educational tools is to disseminate accurate seismic-related knowledge effectively to both the general public and professionals lacking engineering backgrounds (Musacchio et al., 2016). The authors believe that visual content supported by practical models aids non-specialized audiences in understanding fundamental principles governing building behavior during earthquakes (Wang, 2022). By promoting knowledge about seismic design, these videos and models serve as potent tools for raising community awareness about disaster risk reduction (Benadusi, 2014), particularly on those risks associated with constructing or modifying buildings without considering their seismic behavior.

4.1. Importance of Videos' Free Access and Full Availability in Rising Seismic Awareness

Increasing public awareness is crucial for mitigating loss of life and economic damage during earthquakes (Nathe, 2000). Achieving this through educational videos requires two key components. Firstly, videos should be freely accessible, leveraging platforms like YouTube to maximize accessibility across diverse demographics. Secondly, timely availability is critical; viewership statistics (see Table 4) illustrate that views surge following significant earthquakes. This underscores the importance of having educational resources readily available during such critical periods.

4.2. Videos as a Learning Tool

In line with related research highlighting the use of key components to balance cognitive load (de Koning et al., 2009), video-based learning strategies in this study were enhanced by signalling (Ibrahim et al, 2012). As expected, signalling did facilitate students' retention of the subjects (Nevid & Lampmann, 2003), through three strategies:

- (i) introducing a simplified version of the theoretical framework by adding texts, images, and tailored animations (Moreno, 2007);
- (ii) highlighting specific aspects during the explanations by freezing an image or “zooming-in” to focus on specific details (Castro-Alonso et al, 2019); and
- (iii) complementing the actual testing with annotations and legends to facilitate the assimilations of key concepts, measures, or data (Kruger & Doherty, 2016).

Since the videos are always available online, students valued the flexibility to watch content at their convenience and repeatedly if necessary, reinforcing positive attitudes towards this learning method (Kelly et al., 2009). Although videos were kept relatively short—following suggestions in related literature (Ahn & Bir, 2018)—averaging around 7 minutes and 33 seconds, statistical evidence (see Table 4) suggests that even shorter durations could enhance effectiveness, given that average viewing times were approximately 50% of video lengths, suggesting optimal lengths closer to 3 minutes and 45 seconds.

The 50% average watching time refers to views of the videos by random people, including the students who were tasked with fabricating the models presented in them. One may suppose that, contrary to regular viewers, students would tend to watch the whole video. To test this hypothesis, we compared the probable time at which people usually stopped watching the video against the times when the most and least frequently chosen questions were actually answered in the videos (Table 9). Except for video #3, the most frequently chosen questions were

answered within the average watching time. Regarding the least frequently chosen questions, except for videos #1 and #3, all questions were answered beyond the average watching time.

These matching trends seem to indicate that students also watched the videos for an average time equal to half of the actual video length, even when they were supposed to watch it in full. Whether this occurred due to students losing interest in the video or having short attention spans, this observation seems to confirm the idea that shorter videos (approximately 3 minutes and 45 seconds) would be more effective as a learning tool.

Table 9. Comparison of the probable time at which people usually stopped watching the video against the times where the answers to the most and least frequently chosen questions were given.

#	Video	Inferred average time at which people stopped watching the video*	time at which the answer was given in the video	
			for the most-frequently chosen questions**	for the least-frequently chosen questions***
1	Natural Period of Buildings	02:56:35	00:30:00	00:49:00
2	Lateral Force-Resistant Systems	04:20:23	02:31:00	06:00:00
3	Diaphragms and Openings	03:58:17	06:56:00	03:04:00
4	Building Configuration Irregularities - Part 1: Torsion	03:58:41	00:25:00	04:15:00
5	Building Configuration Irregularities - Part 2: Irregular Plans & Pounding	04:34:24	01:32:00	07:19:00
6	Building Configuration Irregularities - Part 3: Short Columns & Soft Storey	03:43:53	01:05:00	04:36:00
7	Non-structural Elements: Infill Walls	02:41:40	02:00:00	02:50:00

* based on Table 4.

,* in reference to questions displayed in Table 5 and Table 6, respectively.

4.3. Monothematic Videos Versus Short Columns & Soft Storey Mechanisms

All videos, except for video # 6—Building Configuration Irregularities - Part 3: Short Columns & Soft Storey, are monothematic. Unlike video # 5, which also includes two concepts (pounding and irregular layouts), the two topics presented in video #6 are not necessarily related nor were explained as such. Hence, in practice, there are two distinct topics within this single video.

The decision of merging them was based on that, otherwise, separate videos would be too short. However, judging by the contradictory results of the surveys, this might not have been the best idea to enhance common knowledge on these subjects. Simultaneously, but not surprisingly, video #6 is the most watched video on YouTube (Table 4).

Upon close examination of the surveys' results, questions related to video #6 display a fairly erratic pattern. Question 34—Why soft storey mechanisms are dangerous?—is one of the least frequently chosen question to answer across all surveys (see Table 6). Conversely, question 29—What is a short column?—is one of the most frequently chosen (see Table 5). This confidence seems to be reflected in the average responses to question 31—Which design solutions help to prevent short columns mechanisms in buildings during earthquakes?—which showed a significant increase between S2-1 and baseline surveys (see Table 7). Although this

increase may indicate gains in knowledge after watching the videos, no question from video #6 was among the highest increases between S3 and S2-2 (see Table 8). Moreover, question 30—In which situations short column mechanisms tend to appear in buildings?—displayed the largest negative difference between validation and key answers in S2-2 (Figure 6) and remained negative even after S3 (Figure 7).

Due to these inconsistencies in the students' responses, whenever mentioned during the interview by a student, we asked them to explain the concepts of soft storey and/or short column mechanisms in their own words. From these explanations, it seems that while the concept of a soft storey is generally fairly understood, the notion of a short column—as suspected, is highly misleading. For example, one student correctly elaborated on the soft storey: “There was an example from [a building in] Türkiye with commercial areas in the ground floors, so without walls. But the upper floors had walls because those were residential. This created the soft storey problem. Also, in another example, [...] there is an apartment [building] with 3 stories, and the middle one is higher than the others, so this also creates the [soft-storey] problem.” (St03).

The confusing aspect of the short column irregularity is that, while it seems like a quite obvious issue, its negative implications for the seismic behaviour of buildings are far less clear: “A short column is a column shorter than the other ones and it affects [the] *statics* [sic] of the building because it doesn't have balance.” (St08). Another student openly expressed that she did not understand this concept and when asked why, she answered: “Because I didn't see it or hear it before. For example, where can we use that kind of columns?” (St01).

Another, or perhaps additional, explanation could be rooted in an oversimplification of the phenomena behind the soft storey and short columns mechanisms. Since the use of educational videos is somewhat limited to basic and simplified language, this simplification might undermine the efforts to introduce rather elaborated analyses or notions. In the pursuit of a wider overview rather than an in-depth understanding, such simplified explanations often avoid the inherent complexity of the dynamics involved in the seismic response of real structures, such as multi-storey buildings. For this reason, to prevent misunderstanding due to oversimplification, basic videos could be complemented with media targeting advance knowledge and a highest level of detailing.

4.4. Learning by Making The Models or Testing The Models?

If we accept that positive variations in the average scores of successive surveys indicate students' increased learning, then the major knowledge gains took place after students watched the videos. Similarly, the comparatively small increases in S3 with respect to S2-2 could be understood as if working with models effectively enhanced students' understanding. However, this effect may be influenced, at least partially, by students learning by re-watching the videos for the purpose of making the models.

Despite the predominance of video-based learning, the survey results provide some evidence that students learned from the working with models, especially in video # 4. Question 22 about eccentricity appears as one of the least frequently selected, but then it also appears in the largest increase towards S3. This is a clear indication that students who particularly worked with those models did increase their understanding of the video subjects by making and testing the models. Similarly, the larger increase in S3 with respect to S2-2 of question 23 seems to prove that by making models of video #4, students did improve their understanding of buildings' torsional behaviour.

The idea that videos might have been more effective as a learning tool than models is reinforced by the interview results. When forced to choose between videos or models to hypothetically teach the survey contents, the overall opinion is that most of the survey contents could be

explained using only videos. Moreover, during the interviews, the unanimous opinion was that videos were useful for making the models.

Another overall opinion is that some topics of videos #3 and #7 could be explained by models. The students who worked with videos #1 and #2 believed that at least two questions in there could be explained by models. Interestingly, the main issue here is that “teaching with models” does not explicitly refer to either making or testing them, so it is unclear what this choice means in the minds of the students.

During the interviews, students expressed that they learned by *testing* the models, yet they seem to see little-to-no knowledge gain in *making* the models. This may appear contradictory, yet one way to interpret it is that students value the fact that models (as used in the videos) do help them to grasp the targeted learning outcomes, but they tend to dismiss the making of the model as a learning tool for seismic design inputs.

The interviews showed that the most probably reason for model making not significantly helping students to increase their knowledge is that they focused more on the fabrication aspects of the models, rather than on understanding the phenomena these were intended to demonstrate.

Models, unexpectedly, posed a great challenge for the students in terms of fabrication. During the interviews, most of their observations focused on how easy or difficult it was to make the assigned model and how much information they could obtain from the videos. They even tended to judge the videos based on this latter perspective. One plausible explanation for this struggle is the fact that these students had their first year of architecture under an online education system—thanks to COVID-19—so by the time they were asked to make the models, they had only one semester of experience in model making.

Despite the models being supposed to reinforce the lessons learned in the videos, they became an independent problem to be solved in a short time (one week). Therefore, the effects of the simulated earthquake shaking were important for the students only in terms of whether the model worked as in the video or simply collapsed.

4.5. Surveys’ Reliability

The relatively small variations between the average scores of validation survey and key answers are a clear sign of the reliability of the assessment. However, despite a good level of agreement between answers (students and key), the positive difference in S2-2 seems to indicate hesitation of the students, while the appearance of negative variations in S3 may indicate overconfidence of students in certain aspects. This overconfidence might be natural or perhaps is the result of students being exposed constantly to the same subject for three weeks. This cannot be attributable to low-skilled students typically overestimating their performance—the Dunning-Kruger effect (Feld et al., 2017)—simply because if so, overconfidence would have been evident also in S2-2. Among several factors that could possibly produce overconfidence, including gender, cultural background, educational levels, and even performing hard tasks, one feasible factor in this case may be overconfidence due to having more information (Oskamp, 1965—as cited in (Wüst & Beck, 2018)).

5. CONCLUSION

The presented study aimed to increase seismic-related knowledge of non-specialized audiences using videos and physical models. The freely available videos on YouTube, coupled with enhanced signalling, are the key features that make these educational media effective in raising seismic awareness in communities susceptible to be affected by future earthquakes.

A subsequent workshop with architecture students was conducted to assess the effectiveness of this method as a learning tool. Post-workshop surveys and interviews with students revealed the following key findings:

- Videos produced in the presented way are effective tools for enhancing learning in students without prior knowledge of seismic design of buildings.
- Educational media dealing with complex subjects, such as seismic design, should be monothematic. This study suggests that merging two subjects within one video prevent students from a clear understanding of key concepts.
- Evidence from surveys and interviews indicates that students attribute knowledge gains to models when used to explain specific topics, such torsional behaviour, the role of diaphragms, and performance of non-structural components.
- Despite the positive evaluation of *testing* models, students, in general, perceive little-to-no knowledge gain in *making* the models. The reasons for this perception are uncertain; however, one plausible argument is that the participant cohort had little experience working with physical models.

These observations may serve as guidelines for incorporating these tools into teaching strategies for seismic-related courses in Architecture.

Finally, given the necessary simplified language and narrative of these educational videos, complex phenomena associated with the actual behaviour of buildings under earthquakes might have been overlooked. Therefore, future work should complement these basic videos with media targeting advance knowledge and a higher level of detailing.

Acknowledgments

The authors would like to thank Merve Hilal Aktaş, Halis Arda Ozdemir, and Kutay Altunkaynak for their contribution to the production of models and videos described in the study. This study was partially funded by Yasar University Project Evaluation Commission (PDK) with the project BAP103 ‘Deprem ve Binalar’. A preliminary version of this study paper was presented in the Sixth International Conference on Earthquake Engineering and Seismology, held in Gebze, Kocaeli, Turkey, in October 13-15, 2021.

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:** Yaşar University Ethics Commission, 31.01.2023-03.

Contribution of Authors

Mauricio Morales-Beltran: Conception, Resources, Methodology, Design and Visualization, Supervision, Materials, Data collection and Processing, Analysis, Writing. **Ecenur Kızılörenli:** Resources, Design and Visualization, Supervision, Materials, Data collection and Processing, Literature Review, Writing. **Ceren Duyal:** Design, Materials, Data collection, Literature Review, Writing.

Orcid

Mauricio Morales-Beltran  <https://orcid.org/0000-0003-4883-4314>

Ecenur Kızılörenli  <https://orcid.org/0000-0002-3992-1363>

Ceren Duyal  <https://orcid.org/0000-0002-5229-5299>

REFERENCES

- Adeoye-Olatunde, O.A., & Olenik, N.L. (2021) Research and scholarly methods: Semi-structured interviews. *Journal of the American College of Clinical Pharmacy*, 4, 1358–1367. <https://doi.org/10.1002/jac5.1441>
- Ahn, B., & Bir, D.D. (2018). Student Interactions with Online Videos in a Large Hybrid

- Mechanics of Materials Course. *Advances in Engineering Education*, 6(3), 1-24
- BAP103-Deprem & Binalar. (2021). *Earthquakes & Buildings*. YouTube. <https://www.youtube.com/@earthquakesbuildings5063>
- Benadusi, M. (2014). Pedagogies of the unknown: Unpacking ‘culture’ in disaster risk reduction education. *Journal of Contingencies and Crisis Management*, 22(3), 174-183.
- Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University* (4th ed.). McGraw-hill education (UK).
- Binici, B., Yakut, A., Canbay, E., Akpınar, U., & Tuncay, K. (2022). Identifying buildings with high collapse risk based on samos earthquake damage inventory in İzmir. *Bulletin of Earthquake Engineering*. <https://doi.org/10.1007/s10518-021-01289-5>
- Blackmore, K., Compston, P., Kane, L., Quinn, D., & Cropley, D. (2010). *The Engineering Hubs and Spokes Project-institutional cooperation in educational design and delivery*. University of Queensland.
- Brame, C. J. (2016). Effective educational videos: principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education*, 15(4), es6. <https://doi.org/10.1187/cbe.16-03-0125>
- Bravo, E., Amante, B., Simo, P., Enache, M., & Fernandez, V. (2011). Video as a new teaching tool to increase student motivation. *2011 IEEE Global Engineering Education Conference (EDUCON)*, 638–642. <https://doi.org/10.1109/EDUCON.2011.5773205>
- Bregger, Y.A. (2017). Blended learning: Architectural design studio experiences using housing in Istanbul. *Journal of Problem Based Learning in Higher Education*, 5(1), 126-137.
- Castro-Alonso, J. C., Ayres, P., & Sweller, J. (2019). Instructional visualizations, cognitive load theory, and visuospatial processing. *Visuospatial Processing for Education in Health and Natural Sciences*, 111-143. https://doi.org/10.1007/978-3-030-20969-8_5
- Charleson, A.W. (2018). Earthquake engineering education in schools of architecture: developments during the last ten years including rule-of-thumb software. *Journal of Architectural Engineering*, 24(3), 4018020, 1-7. [https://doi.org/10.1061/\(ASCE\)AE.1943-5568.0000324](https://doi.org/10.1061/(ASCE)AE.1943-5568.0000324)
- de Koning, B.B., Tabbers, H.K., Rikers, R.M.J.P., & Paas, F. (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review*, 21(2), 113–140. <https://doi.org/10.1007/s10648-009-9098-7>
- Dener, A. (1994). The effect of popular culture on urban form in Istanbul. In *The Urban Experience: A People-Environment Perspective*, London: E. & F. Spon.
- DiCicco-Bloom, B. and Crabtree, B.F. (2006). The qualitative research interview. *Medical Education*, 40, 314-321. <https://doi.org/10.1111/j.1365-2929.2006.02418.x>
- Feld, J., Sauermann, J., & de Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*, 68, 18–24. <https://doi.org/10.1016/j.socec.2017.03.002>
- Green, R.A. (2008). Unauthorised development and seismic hazard vulnerability: A study of squatters and engineers in Istanbul, Turkey. *Disasters*, 32(3), 358-376. <https://doi.org/10.1111/j.1467-7717.2008.01044.x>
- Gulkan, P., Aschheim, M., & Spence, R. (2002). Reinforced concrete frame building with masonry infills. In *World Housing Encyclopedia, Housing report* (Vol. 64).
- Gunasagaran, S., Mari, M.T., Kuppusamy, S., Srirangam, S., & Mohamed, M.R. (2021). Learning construction through model making and its application in architecture design studio. *International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies*, 12(11), 1–10.
- Hajhashemi, K., Caltabiano, N., & Anderson, N. (2016). Students’ perceptions and experiences towards the educational value of online videos. *Australian Educational Computing*,

- 31(2).
- Hussain, E., Kalaycıoğlu, S., Milliner, C.W.D., & Çakir, Z. (2023). Preconditioning the 2023 Kahramanmaraş (Türkiye) earthquake disaster. *Nature Reviews Earth & Environment*, 4(5), 287–289. <https://doi.org/10.1038/s43017-023-00411-2>
- Iban, M.C. (2020). Lessons from approaches to informal housing and non-compliant development in Turkey: An in-depth policy analysis with a historical framework. *Land Use Policy*, 99, 105104. <https://doi.org/10.1016/j.landusepol.2020.105104>
- Ibrahim, M., Antonenko, P.D., Greenwood, C.M., & Wheeler, D. (2012). Effects of segmenting, signalling, and weeding on learning from educational video. *Learning, Media and Technology*, 37(3), 220-235. <https://doi.org/10.1080/17439884.2011.585993>
- Iskander, M. (2007). *Innovations in E-learning, instruction technology, assessment and engineering education*. Springer Science & Business Media.
- Ji, T., & Bell, A. (2000). *Seeing and touching structural concepts in class teaching*. 26–28.
- Kallio H., Pietilä A.-M., Johnson M. & Kangasniemi M. (2016) Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72(12), 2954–2965. <https://doi.org/10.1111/jan.13031>
- Kelly, M., Lyng, C., McGrath, M., & Cannon, G. (2009). A multi-method study to determine the effectiveness of, and student attitudes to, online instructional videos for teaching clinical nursing skills. *Nurse Education Today*, 29(3), 292-300. <https://doi.org/10.1016/j.nedt.2008.09.004>
- Kruger, J.L., & Doherty, S. (2016). Measuring cognitive load in the presence of educational video: Towards a multimodal methodology. *Australasian Journal of Educational Technology*, 32(6). <https://doi.org/10.14742/ajet.3084>
- López, D.L., Rodríguez, M.D., & Costas, S.G. (2022). *Intuition and experimentation as teaching tools: Physical and interactive computational models*. 9727–9734.
- Morales-Beltran, M., Kızılörenli, E., Duyal, C., Aktaş, M., Ozdemir, H., & Altunkaynak, K. (2021). *Deprem ve Binalar: Eğitsel medya kullanımı ile deprem ve binaların sismik davranışı hakkındaki temel bilgilerin halka sağlanması [Earthquake and Buildings: Providing citizens with basic knowledge on the seismic behaviour of buildings using educational media]*. 6th International Conference on Earthquake Engineering and Seismology (6ICEES), Gebze, Türkiye.
- Morales-Beltran, M., & Yildiz, B. (2020). Integrating configuration-based seismic design principles into architectural education: Teaching strategies for lecture courses. *Architectural Engineering and Design Management*, 1-19. <https://doi.org/10.1080/17452007.2020.1738995>
- Moreno, R. (2007). Optimising learning from animations by minimising cognitive load: Cognitive and affective consequences of signalling and segmentation methods. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 21(6), 765-781. <https://doi.org/10.1002/acp.1348>
- Musacchio, G., Falsaperla, S., Sansivero, F., Ferreira, M.A., Oliveira, C.S., Nave, R., & Zonno, G. (2016). Dissemination strategies to instil a culture of safety on earthquake hazard and risk. *Bulletin of Earthquake Engineering*, 14, 2087-2103. <https://doi.org/10.1007/s10518-015-9782-6>
- Napakan, W., Gu, N., Gul, L., & Williams, A. (2009). *Nu Genesis: A Journal of Unique Designs in a Virtual Collaborative Design Studio*, 257–265.
- Nathe, S.K. (2000). Public education for earthquake hazards. *Natural Hazards Review*, 1(4), 191-196. [https://doi.org/10.1061/\(ASCE\)1527-6988\(2000\)1:4\(191\)](https://doi.org/10.1061/(ASCE)1527-6988(2000)1:4(191))
- Nevid, J.S., & Lampmann, J.L. (2003). Effects on content acquisition of signaling key concepts in text material. *Teaching of Psychology*, 30(3), 227-230. https://doi.org/10.1207/S15328023TOP3003_06

- Nuhfer, E., & Knipp, D. (2003). 4: The Knowledge Survey: A Tool for All Reasons. *To Improve the Academy*, 21(1), 59–78. <https://doi.org/10.1002/j.2334-4822.2003.tb00381.x>
- Özmen, C., & Ünay, A.İ. (2007). Commonly encountered seismic design faults due to the architectural design of residential buildings in Turkey. *Building and Environment*, 42(3), 1406–1416. <https://doi.org/10.1016/j.buildenv.2005.09.029>
- Partridge, H., Ponting, D., & McCay, M. (2011). *Good practice report: Blended learning*. Australian Learning and Teaching Council.
- Simonacci, V., & Gallo, M. (2017). Statistical tools for student evaluation of academic educational quality. *Quality & Quantity*, 51(2), 565-579. <https://doi.org/10.1007/s11135-016-0425-z>
- Spence, R. (2004). Risk and regulation: Can improved government action reduce the impacts of natural disasters? *Building Research & Information*, 32(5), 391-402. <https://doi.org/10.1080/0961321042000221043>
- Wang, N. (2022). *Effective Video Solutions for Earth Science Education* (Doctoral dissertation). University of Texas at Dallas, USA
- Watson, J. (2008). *Blended Learning: The convergence of online and face-to-face education. promising practices in online learning*. North American Council for Online Learning.
- Wirth, K.R., & Perkins, D. (2005). Knowledge surveys: An indispensable course design and assessment tool. *Innovations in the Scholarship of Teaching and Learning*, 1–12.
- Wüst, K., & Beck, H. (2018). “I Thought I Did Much Better”-Overconfidence in University Exams. *Decision Sciences Journal of Innovative Education*, 16(4), 310-333. <https://doi.org/10.1111/dsji.12165>
- Yakut, A., Sucuoğlu, H., Binici, B., Canbay, E., Donmez, C., İlki, A., Caner, A., Celik, O.C., & Ay, B.Ö. (2022). Performance of structures in İzmir after the Samos island earthquake. *Bulletin of Earthquake Engineering*, 20(14), 7793-7818. <https://doi.org/10.1007/s10518-021-01226-6>

APPENDICES

Appendix 1: Full Survey Questionnaire

Table 10. Full questionnaire used for knowledge and validation surveys

Video	Question
Video # 1: Natural Period of Buildings	1 What is an earthquake?
	2 What are seismic waves?
	3 What are ground motions?
	4 How ground motions affect buildings?
	5 What is the natural period of a building?
	6 What are the factors that influence the natural period of a building?
Video # 2: Lateral Force-Resistant Systems	7 Why vertical continuity is fundamental to provide buildings with adequate resistance to earthquakes?
	8 Why seismic resistance must be provided in both orthogonal plan directions?
	9 What are the most common seismic-resisting systems used in buildings?
	10 What is the most effective seismic-resisting system?
	11 What is the least effective seismic-resisting system?
	12 What are the most common configurations of braced frames?
Video # 3: Diaphragms and Openings	13 Which structural elements in buildings are considered as diaphragms?
	14 What is the role of a diaphragm in providing seismic resistance to buildings?
	15 Why openings might jeopardize the structural integrity of a diaphragm?
	16 From a seismic-resistance perspective, where is the worst location to make openings in the diaphragm?
	17 From a seismic-resistance perspective, where is the best location to make openings in the diaphragm?
	18 Why discontinuity of the diaphragm might lead to excessive deformations in the structure during earthquakes?
Video # 4: Building Configuration Irregularities - Part 1: Torsion	19 What is Centre of Mass?
	20 What is Stiffness?
	21 What is Centre of Resistance?
	22 What is Eccentricity?
	23 How does eccentricity affect the torsional behaviour of a building during an earthquake?
	24 Why placing the seismic-resistant elements symmetrically in plan is the best way to avoid building torsion?
Video # 5: Building Configuration Irregularities - Part 2: Irregular Plans & Pounding	25 What are regular and irregular building plan layouts?
	26 Why irregular plan layouts can be potentially dangerous during earthquakes?
	27 What is the main benefit of separating volumes of irregular building configurations using a gap?
	28 In practice, how wide should the seismic gap be?
Video # 6: Building Configuration Irregularities - Part 3: Short Columns & Soft Storey	29 What is a short column?
	30 In which situations short column mechanisms tend to appear in buildings?
	31 What design solutions help to prevent short columns mechanisms in buildings during earthquakes?
	32 Why short column mechanisms are dangerous?

	33	What is a soft storey?
	34	Why soft storey mechanisms are dangerous?
	35	Why in Türkiye buildings with soft storeys are very common?
Video # 7: Non-structural Elements - Infill Walls	36	What is the most common type of non-structural infill wall used in Türkiye?
	37	What are the negative effects of using infill walls directly connected to the structural frames?
	38	What type of damage appears when infill walls resist in-plane inertia forces?
	39	What type of damage may appear when infill walls are subjected to out-of-plane forces?
	40	Why separating infill walls from the structural frames by a gap significantly reduces the possibility of non-structural damage during an earthquake?

Appendix 2: Most frequent question-to-be-answered per video

Table 11. Most frequent question-to-be-answered per video, in each survey

question	survey	%	all answers			only tested answers			key answers			delta		
			\bar{x}	M_d	M_o	\bar{x}	M_d	M_o	\bar{x}	M_d	M_o	\bar{x}	M_d	M_o
Q1	S1		2.75	3	3	-	-	-	-	-	-	-	-	-
	S2-1		2.71	3	3	-	-	-	-	-	-	-	-	-
	S2-2	63	2.73	3	3	2.71	3	3	2.54	3	3	-0.17	0	0
	S3	82	2.69	3	3	2.70	3	3	2.75	3	3	0.05	0	0
Q7	S1		1.75	2	2	-	-	-	-	-	-	-	-	-
	S2-1		2.08	2	2	-	-	-	-	-	-	-	-	-
	S2-2	37	2.24	2	3	2.47	3	3	2.16	2	2	-0.32	0	0
	S3	33	2.10	2	2	2.33	2	3	2.27	2	3	-0.15	0	0
Q17	S1		1.18	1	1	-	-	-	-	-	-	-	-	-
	S2-1		2.24	2	3	-	-	-	-	-	-	-	-	-
	S2-2	25	2.22	2	2	2.62	3	3	2.23	2	3	-0.38	0	0
	S3	27	2.31	2	3	2.67	3	3	2.67	3	3	0.20	0	0
Q19	S1		2.49	3	3	-	-	-	-	-	-	-	-	-
	S2-1		2.65	3	3	-	-	-	-	-	-	-	-	-
	S2-2	64	2.61	3	3	2.59	3	3	2.41	3	3	-0.18	0	0
	S3	60	2.51	3	3	2.52	3	3	2.11	2	3	-0.41	0	0
Q26	S1		1.90	2	2	-	-	-	-	-	-	-	-	-
	S2-1		2.41	3	3	-	-	-	-	-	-	-	-	-
	S2-2	43	2.37	2	3	2.32	2	2	2.00	2	1	-0.41	0	-1
	S3	40	2.43	3	3	2.47	3	3	2.21	3	3	-0.25	0	0
Q29	S1		1.86	2	2	-	-	-	-	-	-	-	-	-
	S2-1		2.45	3	3	-	-	-	-	-	-	-	-	-
	S2-2	46	2.61	3	3	2.61	3	3	2.22	2	3	-0.24	0	0
	S3	38	2.55	3	3	2.67	3	3	1.89	2	2	-0.76	-1	-1
Q36	S1		2.06	2	2	-	-	-	-	-	-	-	-	-
	S2-1		2.22	2	2	-	-	-	-	-	-	-	-	-
	S2-2	53	2.37	3	3	2.64	3	3	2.46	3	3	-0.26	0	0
	S3	43	2.31	2	3	2.63	3	3	2.68	3	3	-0.07	0	0

Appendix 3: Interview Questions

During the past statics course, assignment no. 5, you and your teammates were requested to watch a video, prepare physical models and then make a presentation with them in class:

Video

- Do you remember which video was it?
- can you describe it?
- what was the main goal of that video?
- how long was that video?
- was the video spoken in English or Turkish?
- did this help or not?
- Was there anything special or important in the video?
- Was there anything that drew your attention?
- Or something you really liked about the video?
- Something you disliked?

Working

- how was the experience of working with that video as source of information/inspiration to prepare the models?
- could you understand how to build your model from this and/or other videos?
- If yes, which part of the video helped you the most to make the model?
- If not, how did you figure it out the construction/assemble of the model?
- Was it easy or difficult to make the model?
- Why?

Testing

- Could you and your teammates repeat the behaviour of the model as showed in the videos?
- were you able to explain that behaviour?
- Which part of the testing/presentation was very close to the way is presented in the video?
- was there something missing, e.g. something that appeared in the video but was not repeated in the presentation?
- Was there something that went wrong during your presentation/testing?

Learning

- do you feel you learned something during this exercise/process?
- if so, what?
- did you learn more from watching the videos, making the model, or testing the model? - answer from more to less
- what things you could not understand/learn?
- (only if previous question is answered) What was the main obstacle for you to understand/learn that?

Suggestion

if the exercise is repeated in the future:

- would you suggest to focus only on videos, working only with models, or a combination of both?
- what would you repeat?
- what would you do differently?

Learning Process

- Do you think this exercise contributed to your knowledge about earthquakes?
- If the information conveyed through videos was explained not in videos but with physical lectures, which one do you think would attract more your attention? Why?
- Did the colours used in the videos, the use of signs in certain places and the highlighted information affect your focus on the subject? How? Did they affect your understanding of the subject? How? Can you provide an example of this?
- As you may have noticed, long topics (e.g. building configuration irregularities) were subdivided in two or three short videos. Did this make it easier for you to follow and understand the information described in there?
- Was it helpful for you to remember/learn the knowledge by watching the videos over and over in the model making tasks given after the videos? Did it contribute to the model making process? Did it provide you flexibility for your working pattern?

Final Question

Is there anything you would like to add?

Algebraic knowledge for teaching test: An adaptation study

Ali Bozkurt^{1*}, Begüm Özmuşul²

¹Gaziantep University, Gaziantep Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

²Gaziantep University, Nizip Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

ARTICLE HISTORY

Received: Nov. 05, 2023

Accepted: July 15, 2024

Keywords:

Algebra teaching,
Algebraic knowledge,
Adaptation study,
Preservice Teachers,
MKT-PFA.

Abstract: In this study, the Mathematical Knowledge for Teaching-Elementary Patterns Functions and Algebra-Content Knowledge (MKT-PFA) test, originally developed in English as part of the "Learning Mathematics for Teaching Project" at Michigan University, was adapted into Turkish. The test comprises two equivalent forms, A and B, each translated into Turkish and culturally adapted through consultations with two mathematics education academics and five secondary school math teachers pursuing doctoral studies. A total of 328 pre-service teachers at a Turkish public university's elementary school mathematics teaching department were administered form A (14 questions, 29 items) and form B (12 questions, 27 items) at a one-week interval. Psychometric analyses revealed high reliability (KR-20: A=0.712, B=0.735; Lord reliability: A=0.733, B=0.756), and strong correlations (r_{pti}) with the original English forms, indicating suitable adaptation. Item difficulties analyzed using a one-parameter Item Response Theory model showed a normal distribution, affirming the tests' validity for assessing pre-service teachers' algebra teaching knowledge in Türkiye.

1. INTRODUCTION

Mathematics education is a field that requires interaction between teachers and students in classrooms, professional knowledge, and reasoning to invite students to the learning process of mathematics (Ball et al., 2008). This teaching process, which consists of interactions between teachers and students, helps students act as critical thinkers and develop their reasoning (Cohen, 2011). The teacher's interactions with students in the classroom begin and are developed through the "teaching job". This "teaching job" allows students to reason, interpret, criticize textbook practices on specific topics, use representations correctly, and create examples of mathematical concepts, algorithms, or proofs (Hill et al., 2005). Therefore, teachers should possess certain competencies, such as mathematical knowledge for teaching, interactions with students, technology integration, and understanding of student diversity, to structure their mathematics instruction effectively (Ball et al., 2005; Ma, 1999). These competencies can equip mathematics teachers with the essential skills and knowledge required to enhance their students' achievement and foster positive attitudes toward mathematics. This study focuses on

*CONTACT: Ali BOZKURT ✉ alibozkurt@gantep.edu.tr 📧 Gaziantep University, Gaziantep Faculty of Education, Department of Educational Sciences, Gaziantep, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

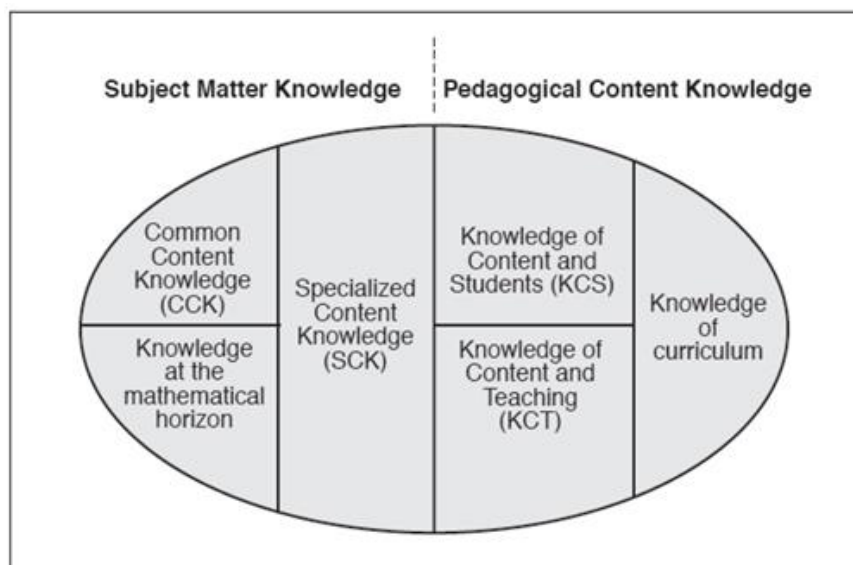
Mathematical Knowledge for Teaching (MKT), one of the competencies identified in the literature, for teachers and preservice teachers to structure the 'teaching job'.

The theoretical basis of mathematical knowledge for teaching (MKT) in instruction is grounded in the idea that what teachers need to know is determined by what teachers do in their teaching practice (Ball, 1990). Studies in the literature (An et al., 2004; Ma, 1999) draw attention to the quantity and quality of teachers' mathematical knowledge for teaching. Additionally, these studies have identified significant differences in mathematics teaching across different countries. This situation is crucial in understanding how teachers acquire mathematical knowledge and apply it in teaching mathematics in varying countries (Ball & Hill, 2008; Hill et al., 2005). MKT assists us in comparing teaching and learning processes across countries (Knipping, 2003). This study presents an adaptation study of the algebra teaching knowledge test for measuring the algebra knowledge for teaching future teachers.

1.1. Mathematical Knowledge for Teaching (MKT)

The knowledge that a teacher should have in the teaching process can be classified under two main headings: Pedagogical content knowledge and subject matter knowledge (Shulman, 1986). The former is concerned with presenting the knowledge in the relevant field to the student by transforming it into a teachable structure while the latter is the knowledge about the basic principles, concepts, laws, and theories stipulated by the curriculum of the field. The components of these types of information are given in Figure 1.

Figure 1. Domain of MKT (Ball et al., 2008, p.403; Shulman, 1986).



In Figure 1, there are three sub-fields under the title of pedagogical content knowledge; namely, "knowledge of content and teaching, knowledge of content and students, and knowledge of content and curriculum" (Ball et al., 2008). Under the title of subject area information, there are three sub-fields: "common content knowledge, specialized content knowledge, and horizon content knowledge". Although this division can be used to analyze subfields, it is intertwined with teaching practices (Kim, 2016; Koellner et al., 2007).

The details of the teaching information vary according to the course. Mathematical knowledge for teaching comes to the fore for the mathematics course (Hill et al., 2005). This information is essential for realizing mathematics teaching (National Mathematics Advisory Panel, 2008). However, the required pedagogical information can vary according to the learning objectives of the mathematics course. One of the learning areas in mathematics is algebra, a field of mathematics that involves developing rules to represent functional expressions and relations, expressing these rules with symbols, writing and solving equations, and making generalizations

from calculations with numbers (Lew, 2004; Welder & Simonsen, 2011). Some researchers (Kieran et al., 2018; Lew, 2004) focus on the abstract features that distinguish it from arithmetic in the definition of algebra and define algebraic thinking as "the ability to operate on an unknown quantity as if the quantity is known, as opposed to arithmetic reasoning involving operations on known quantities" (Langrall & Swafford, 1997, p. 2). Some others (Driscoll, 1999; Zazkis & Liljedahl, 2002) have noted the critical importance of functions that play in algebra, which is characterized as the capacity to represent quantitative situations in algebraic thinking. In both cases, they are part of algebraic thinking which is aimed to be improved in algebra teaching. In the algebra teaching process, students are expected to be taught algebra and gain algebraic thinking skills (Schmittau, 2005). Charalambous (2008) concluded that there is a potent relationship between teacher knowledge and teaching performance. Therefore, the fact that teachers need to know how to teach the basic concepts of algebra may cause students to have difficulty learning algebra. For this reason, Hill and Ball (2009) developed the 'patterns, functions, and algebra' test to measure teachers' and preservice teachers' algebra knowledge for teaching. Within the scope of this study, the adaptation of the test developed by Hill and Ball (2009) was carried out in Turkish. Thus, with the adaptation of this achievement test, the algebra knowledge for teaching preservice teachers can be assessed. Consequently, based on the levels of algebra knowledge for teaching among preservice teachers, mathematics educators can enrich the scope of the algebra teaching course as specified by the Council of Higher Education (CoHE (YÖK: Yükseköğretim Kurumu), 2018). Teachers go through the candidacy process to gain competence in the professional context. Morris et al. (2009) mentioned that preservice teachers could define mathematical concepts but could not spontaneously apply planning or assessment of teaching and learning in line with their learning objectives. Huang and Kulm (2012) indicated that preservice teachers need more knowledge about the place of the term of function in the curriculum, its teaching, and content knowledge. In particular, the study concluded that the flexibility of the preservice teachers in the use of different representations and the weakness in the selection of function perspectives. He also recommended that the teacher training program should provide content areas that are consistent with the curriculum. Strand and Mills (2014) stated that preservice teachers used the "guess and check" strategy while using variables to represent unknown numbers in algebra problems and then writing numbers instead of variables while solving. Thus, preservice teachers are in different thinking processes to confirm their ideas.

1.2. Mathematics Teacher Education Program in Türkiye

The General Competencies for the Teaching Profession, which outline the knowledge, skills, and attitudes necessary for effectively and efficiently fulfilling the teaching profession, were updated in 2017. In addition, the Teacher Strategy Paper was published in 2017. In the mentioned documents, new goals and expectations, as well as new competencies related to teaching, are included. In addition, some official documents such as the 10th Development Plan (2014-2018), the Strategic Plan of the Ministry of National Education (2015-2019), Türkiye's higher education qualifications framework, educational sciences field qualifications, and teacher training have been published over time (CoHE (YÖK), 2018). Considering the developments required in teacher training, as well as the structural changes in the Turkish education system, societal demands, and social needs, the necessity of updating teacher education undergraduate programs has emerged. (CoHE (YÖK), 2018). In this direction, the elementary education mathematics teacher undergraduate program was changed in 2018. While field courses such as algebra, differential equations, and elementary number theory were intense in the curriculum before 2018, since 2018, mathematics education, such as teaching algebra, geometry and measurement, numbers, statistics, and probability has begun to be given more place.

To determine the mathematics teaching knowledge of teachers and preservice teachers, tests are developed specifically for various learning areas. However, using these tests directly to assess the situation in different countries may not yield reliable results. In this context, it is

necessary to adapt the developed measurement tool for each country to be applied, so much so that many studies have revealed that the characteristic features of the teaching systems they examine are influenced by culturally located teaching practices (Delaney et al., 2008; Knipping, 2003; Ma, 1999; Stiegler & Hiebert, 1999; Wilson et al., 2001).

In the field of mathematics education, the Learning Mathematics for Teaching (LMT) project has developed MKT tests to measure teachers' knowledge of mathematics teaching (Ball & Hill, 2008; Hill & Ball, 2004; Hill & Ball, 2009). Adaptation studies of MKT tests developed in the USA to different languages and cultures were carried out. Some of these are the following:

- Delaney et al. (2008) adapted the forms developed for the learning domains of numbers and operations, algebra, and geometry from MKT tests for use in Ireland. They found that some Irish teachers were unsure of the meaning of certain terms and suggested changes to the general cultural context for adaptation.
- Mosvold and Fauskanger (2009) determined that there was a need for significant changes in the process of adapting the form developed for the geometry learning domain from the MKT scales to Norway. For example, it has been observed that some concepts in the scale are not found in the Norwegian curriculum. Some changes have been made to make it more usable, valid, and reliable for Norwegian teachers.
- Ng et al. (2012) found some contextual problems and differences in teaching practices and representations in the process of adapting the form developed for the geometry learning field, one of the MKT scales, to Indonesia.
- Cole (2012) found cultural incompatibility between America and Ghana in the questions in the form developed for the learning domain of numbers and operations from MKT scales.
- Kim (2020) conducted a study on adapting the form developed for the algebra learning field from the MKT scales. In the study, it was determined that Korean teachers had a high rate of answering the MKT test correctly, but the relationship between teaching methods and algebraic reasoning was low.
- Esendemir and Bindak (2019) adapted the geometry teaching knowledge scale, which is a learning area of mathematics, of secondary school mathematics teachers into Turkish.

When the studies on the adaptation of MKT in the literature above are examined, it is seen that there are fewer adaptation studies of forms measuring algebraic knowledge for teaching. Regarding the field of algebra learning, Delaney et al. (2008) observed that Irish teachers examined their algebraic knowledge for teaching and adapted it to their own culture. Similarly, Kim (2020) observed that Korean teachers adapt their algebraic knowledge for teaching to their own culture to measure it. Within the scope of this study, the Mathematical Knowledge for Teaching-Elementary Patterns Functions and Algebra Content Knowledge (MKT-PFA) forms given in Hill and Ball (2009) were adapted to measure the algebra teaching knowledge of pre-service teachers. When the literature is examined, there is no Turkish adaptation of an algebra teaching tool used internationally. In this respect, these achievement tests measuring algebraic knowledge for teaching preservice mathematics teachers in Türkiye are expected to contribute to the literature.

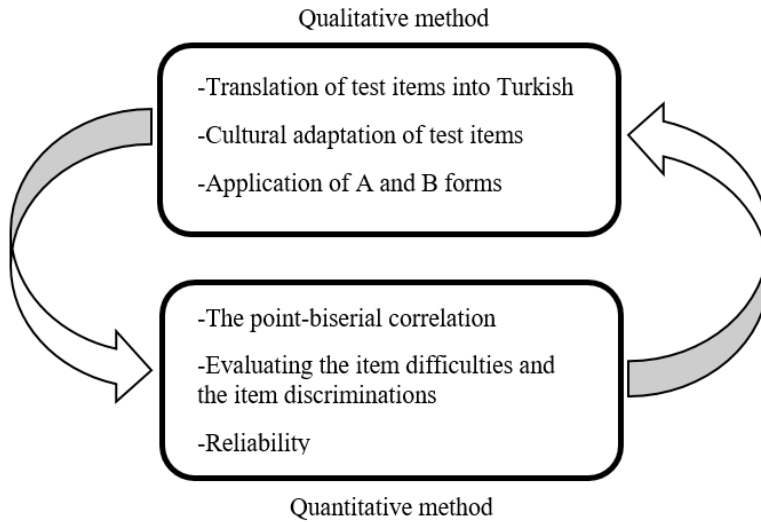
2. METHOD

The study is aimed to adapt the MKT-PFA test which was developed in English by Hill and Ball (2009) within the scope of the "Learning Mathematics for Teaching Project" carried out at the Michigan University, into Turkish. Both qualitative and quantitative methods were used in the adaptation process of the test (see [Figure 2](#)).

Qualitatively, to adapt the test items to Turkish, analysis was conducted as a result of interviews with field experts in line with the adaptation framework in the Delaney et al. (2008) study. The adaptation framework proposed by Delaney et al. (2008), which was used in the present study,

was also used in adaptation studies conducted to measure mathematical knowledge for teaching of teachers and preservice teachers in different countries (Esendemir & Bindak, 2019; Kim, 2020; Mosvold & Fauskanger, 2009; Ng, 2012; Ng et al., 2012). In this respect, this framework was considered to be sufficient. Quantitative data were analyzed using psychometric methods.

Figure 2. *The adaptation process of MKT-PFA.*



2.1. Algebraic Knowledge for Teaching Test

With the A and B equivalent forms of the MKT-PFA test, it was aimed to measure the algebraic teaching knowledge of teachers and preservice teachers in the United States. The tests were developed to examine the structure of teachers' and preservice teachers' teaching knowledge (Hill et al., 2004), how teachers learn to teach mathematical knowledge (Hill & Ball, 2004), and how teacher knowledge is related to achievements in students' mathematics achievement (Ball et al., 2005). There are 14 questions in Form A of the MKT-PFA test. In this form, participants are asked to evaluate the options of the 4th, 9th, 14th, 20th, and 22nd questions within the scope of "yes, no and I'm not sure" options. Form B has 12 questions. In this form, participants are asked to evaluate the options of the 6th, 13th, 16th, 19th, and 25th questions within the scope of the "yes, no and I'm not sure" options.

2.2. Adaptation Process of Algebraic Knowledge for Teaching Test

A review of the related literature reveals that, for the adaptation studies of the scales, (i) the test items should be translated into the language to be adapted, (ii) the items should be culturally adapted, (iii) the test should be applied to the relevant sample group, and (iv) validity and reliability studies should be done (Delaney et al., 2008). In this study, these stages were followed within the scope of adapting the items in the A and B forms of the MKT-PFA test to Turkish to determine the algebra teaching knowledge levels of preservice elementary school mathematics teachers in Türkiye.

2.2.1. Translation of test items into Turkish

In the first stage of the adaptation studies, the items in the A and B forms of the MKT-PFA test were translated from English to Turkish. An English education expert was consulted during the translation process of the test items. The cultural conformity of the terms has not been taken into account when translating the texts. For this reason, it was assumed that the test items did not undergo any changes in this process. Thus, without changing the mathematical substance of the test items, a one-to-one translation was made from English to Turkish.

2.2.2. Cultural adaptation of test items

In adaptation studies, intercalarily to the translation process into another language, the available test should also be culturally adapted. Materials devoid of cultural components may cause participants to focus on another thing (Hambleton, 1994). This distraction may negatively influence the success of the attendees regarding the items (Yen, 1993). In order to determine whether it reflects the situations that would arise in the classrooms in Türkiye, interviews were conducted with a group of 7 participants. Two of the participants were mathematics educators and five were elementary school mathematics teachers with doctoral degrees in mathematics education, and taught algebra and algebra teaching. During the interviews, the items were adapted according to the following four criteria: (i) changes in the general cultural context, (ii) changes in the context of the school culture, (iii) changes in the mathematical structure, and (iv) changes in the language structure (Delaney et. al., 2008). The group discussed the changes to be made to make each element suitable for Turkish culture. Eventually, a final judgment was made for each change.

A critical question appears regarding the adaptability of a test developed in one country to another: To what extent does the test match the algebra knowledge of elementary preservice mathematics teachers in Türkiye, where the test will be adapted? It is thought that the best mathematics educators and experienced mathematics teachers can answer this question. Therefore, at the end of the interviews, this question was asked to the participants as it is, and all participants agreed that each item in the forms was consistent with the content in Türkiye.

2.2.3. Application of A and B forms of the test to elementary mathematics preservice teachers

The sample sizes most frequently used in previous IRT studies were reviewed while deciding on the sample sizes to apply the Algebra Teaching Knowledge test within the scope of the study. Kline (1994) recommends a sample size of one-tenth (ten times as many participants as the number of items). On the other hand, research in the literature (Pekmezci & Avşar, 2021; Şahin & Anıl, 2017; Yang, 2007) states that at least 150 samples can be created in tests with a single parameter and the number of items between 20-30. Additionally, Sheng (2013) stated that as the sample size increases, there is no significant change in model-data fit values under the unidimensional theory. Additionally, AIC is commonly used as an information criterion for statistical model selection (Burnham & Anderson, 2002). Moreover, AIC tends to perform better with smaller sample groups (Boykin et al., 2023). Similarly, it was observed that there was no significant change in the model-data fit values of the adapted test after 300 samples (Pekmezci & Avsar, 2021). For this reason, the sample of the test to be adapted consists of preservice mathematics teachers studying at the faculty of education of a state university in Türkiye. It was applied to a total of 328 3rd and 4th-grade preservice mathematics teachers, 217 of whom were female and 111 of whom were male, taking the algebra teaching course.

2.3. Situation of Satisfying Item Response Theory (IRT) Assumptions

In Item Response Theory (IRT), the ability parameter that defines a respondent is not dependent on a group of test items (Holmes & Brian, 2019). Another feature that is valid for all models of IRT is that they must meet certain assumptions of IRT. The necessity of meeting these assumptions varies according to IRT's models (Reyhanlıoğlu & Doğan, 2020). One-dimensional IRT has two commonly accepted assumptions: unidimensionality and local independence (Baker & Kim, 2017; Edelen & Reeve, 2007).

Unidimensionality recognizes that the achievement test has a single latent ability (Reyhanlıoğlu & Doğan, 2020). What is sufficient and necessary for this assumption to be met is that there is a dominant component or factor that is measured by the test items and affects test performance. This dominant constituent or factor (element) is called the ability measured by the test (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). In addition, when a one-dimensional test is applied to all populations, the conditional distributions obtained from the test results are

expected to be similar (Hambleton & Swaminathan, 1985). Researchers (Aryadoust et al., 2021; Chou & Wang, 2010; Hambleton et al., 1991; Han, 2022) cited many analyses to show that the test is one-dimensional. The main analytical technique is Exploratory Factor Analysis (EFA). Before performing EFA, the KMO (Kaiser-Meyer-Olkin) statistic value was examined. For form A, the KMO value was determined as 0.722 and the Bartlett's sphericity test statistical value was determined as KMO and Bartlett's Test 1655.537 ($sd = 406, p < 0.05$). For form B, the statistical value of Bartlett's sphericity test was determined as 0.716 and KMO and Bartlett's Test was determined as 1597.755 ($sd = 378, p < 0.05$). If the KMO value is greater than 0.60 and the Bartlett test results show a statistically significant difference, it means that the data is suitable for factor analysis (Tabachnick & Fidell, 2012). Considering the KMO value and Bartlett statistics, it can be said that the sample size is suitable for factorization. When the eigenvalues of the factors for form A were examined, 3 factors were seen above 1. However, while the eigenvalue for the first factor (3.020) is almost 3 times the eigenvalue for the second factor (1.099), the eigenvalue for the second factor (1.099) is twice the eigenvalue for the third factor (1.049). When the eigenvalues of the factors for form B were examined, 3 factors above 1 were observed. However, while the eigenvalue for the first factor (3.970) is almost 3 times the eigenvalue for the second factor (1.155), the eigenvalue for the second factor (1.155) is more than the eigenvalue for the third factor (1.132). Lord (1980) states that a single-factor structure may exist in cases where the eigenvalue of the first factor is significantly greater than the second factor and the eigenvalues of the second factor and the third factor are close to each other. Furthermore, when EFA was conducted on both forms, it was observed that the item loadings of the items in the forms were greater than .30. Upon reviewing studies in the literature (Tabachnick & Fidell, 2012), it is seen that this is considered sufficient. It is seen that PCAR (principal component analysis of residuals), one of these analyses, is used by the test developers. One of these analyses, PCAR (Principal Component Analysis of Residuals), appears to be used by test developers. For this reason, PCAR analysis was performed to show that the adaptation of the MKT-PFA test is one-dimensional. PCAR of the adapted test was obtained as 1.2. According to Smith and Miao (1994), since this value is less than 1.4, it indicates that the adapted test may have one-dimensionality. For this reason, the adapted test is one-dimensional. In addition, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are frequently used information criteria in statistical model selection (Boykin et al., 2023). Both AIC and BIC help in model selection by considering the fit and complexity of a model. For a one-parameter model, BIC is lower than AIC because BIC expresses the complexity of the model. In this framework, AIC and BIC values were calculated in both forms. It was observed that the BIC values of both Form A (AIC: 7704.9 BIC: 7484.905) and Form B (AIC: 6442.714 BIC: 6237.891) were lower than the AIC value. Therefore, it can be said that the tests are uniparametric. In addition, the developers of the MKT-PFA test stated that the test is only aimed at the algebra knowledge for teaching of teachers and preservice teachers.

Items are regressed according to the latent variable in the Rasch measurement; for this reason, the fact that unexplained variances in the items are not related to each other is explained by concept of local independence (Borsboom, 2005). Local independence is when individuals' responses to different items of a test are statistically independent or unrelated to each other (Fan & Bond, 2019; Hambleton & Swaminathan, 1985; Wright, 1996; Yen, 1993). However, for the responses to the items to be statistically independent of each other, the ability measured by the test items must be kept constant (Lord & Novick, 1968). Statistics such as Yen's (1993) 3rd quarter are available to provide local independence assumptions. To ensure the assumptions, the answer to a question in the test must not be a clue for the answer to the other question (Borsboom, 2005; Hambleton & Swaminathan, 1985). Reyhanlıoğlu and Doğan (2020) stated that it is sufficient for the measured structure to be one-dimensional to ensure the local

independence assumption. Accordingly, it can be said that the adapted test meets the local independence assumption because it meets the unidimensionality assumption of MKT-PFA.

2.4. Validity and Reliability

Three psychometric analyses are performed to examine the validity and reliability of the adapted version of a test: Comparing the r_{pbi} (r_{pbi} are the correlation coefficients of the items themselves), evaluating the item difficulties and item discrimination, and calculating the reliability of the form (Delaney et al., 2008; Ng, 2012). In this context, the validity and reliability of the adapted test need to be evaluated, entailing a comparison of the r_{pbi} between the USA and Türkiye measurements, evaluation of the item difficulties and item discriminations of the items in the A and B forms of the MKT-PFA test using a one-parameter IRT model, and the evaluation of the MKT-PFA test. The reliability of the A and B forms was calculated. The KR-20 value was calculated within the scope of the reliability of the test.

Point biserial correlation is used to examine how one item relates to all other items (de Ayala, 2013). The higher the point biserial correlation of an item, the stronger the relationship between that item and the measured construct. In other words, the higher the r_{pbi} of an item, the better it can distinguish individuals whose quality under investigation is closer to each other (Delaney et al., 2008; Ng, 2012; Marcinek et al., 2022). In the context of this study, r_{pbi} with high scores indicates that the items can distinguish teachers with closer algebra knowledge.

Negative r_{pbi} values of an item indicate that teachers with higher mathematics teaching knowledge would probably answer this item incorrectly, and the item may not measure the intended construct. Researchers analyzing LMT item properties evaluated all items with negative r_{pbi} values as poorly functioning (Delaney et al., 2008; Esendemir & Bindak, 2019; Fauskanger et al., 2012; Kim, 2020; Kwon et al., 2012; Marcinek & Partová, 2016; Marcinek et al., 2022; Ng, 2012; Ng et al., 2012). In addition, some studies showed scatterplots (Kim, 2020; Kwon et al., 2012), performed a Fisher Z transform on r_{pbi} values to place them on the interval scale (Delaney et al., 2008; Marcinek et al., 2022), identified outliers (Ng, 2012; Ng et al., 2012) and expressed correlations between the r_{pbi} values of items in the USA and those used in their own countries (Delaney et al., 2008; Esendemir & Bindak, 2019; Ng, 2012).

Items with r_{pbi} value of around zero show no relationship between how respondents answered the item and their general mathematics teaching knowledge level. In other words, when we remove such an item from the test, it cannot be said whether the teacher who gave the correct answer was generally more successful than the teacher who gave the wrong answer (Hambleton et al., 1991). Therefore, r_{pbi} predictive was able to examine the difficulty levels and the overall reliability of the items in the context of the relationship between countries. If there is a difference between these items, it means that these items perform differently between cultures (Cronbach & Shavelson, 2004; Delaney et al., 2008).

A one-parameter IRT model was used to calculate the item difficulty values of the test. Depending on the sample size of the data obtained from the pilot study, researchers can use one- or two-parameter IRT models (Delaney et al., 2008; Ng, 2012; Esendemir & Bindak, 2019). When looking at item difficulty, 0 is considered to represent average teacher skill. Items with a difficulty value of less than 0 are considered easier, and items higher than 0 are considered more difficult (Ng, 2012). In addition, the test information curve maximum was generated for each form to examine how useful the measures were. The test information curve provides information on whether the measures were more difficult or less difficult for the average preservice teachers, i.e., whether the measures can discriminate among preservice teachers of different level of abilities (Baker & Kim, 2017).

Finally, after the final version of the PFA test was provided, the reliability of measurements for Form A and Form B, which calculates how consistent respondents' scores are across multiple items or tests, was computed. Test reliability measures the consistency of test takers' scores on

more than one item (Delaney et al., 2008). A widely used reliability measure from classical test theory is the KR-20. KR-20 is reported in the reliability of achievement tests evaluated as 0-1 (Cronbach & Shavelson, 2004). For the reliability of the test, the KR-20 value is expected to be above .70 (Cronbach & Shavelson, 2004). In addition, Lord reliability is included in the reliability of tests graded as 1 and 0 in IRT (Çelen, 2008; Frary, 1989; Özdemir, 2004). Lord reliability is typically calculated based on item parameters obtained from IRT and individuals' responses (Sireci et al., 1991). This measurement is used to assess the internal consistency of a test and indicates the repeatability of an individual's performance on the test. In the literature, it is also expressed as estimated reliability or reliability coefficient (Embretson & Reise, 2013). Estimated reliability is a measure reflecting how accurately a test measures individuals' true abilities. This reliability measure reflects the internal consistency of the test and indicates that the test items measure consistently with each other (Embretson & Reise, 2013). For this reason, Lord reliability, one of the reliabilities of the test's IRT, is also included.

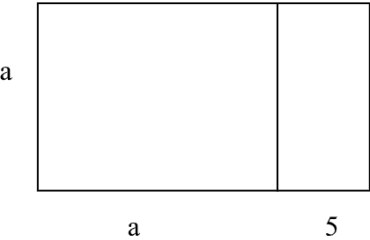
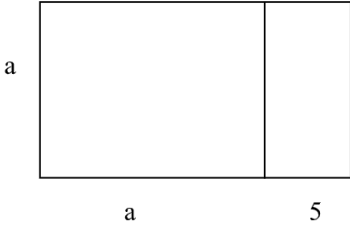
3. RESULTS

In this part of the study, the findings obtained from the cultural adaptation process and psychometric analyses are included to examine the validity and reliability of the Turkish versions of the A and B forms of the MKT-PFA test.

3.1. Cultural Adaptation of Test Items

In the first stage of the adaptation process, the A and B forms of the MKT-PFA test were translated from English to Turkish. In this process, attention was paid to the direct translation of the expressions in the original form. An example of item is given Table 1 in form A of the MKT-PFA test and its translation into Turkish.

Table 1. An item in form A of the MKT-PFA test and its translation into Turkish.

<p>Ms. Whitley was surprised when her students wrote many different expressions to represent the area of the figure below. She wanted to make sure that she did not mark as incorrect any that were actually right. For each of the following expressions, decide whether the expression correctly represents or does not correctly represent the area of the figure. (Mark REPRESENT, DOES NOT REPRESENT, or I'M NOT SURE for each.)</p>	<p>Zeynep öğretmenin öğrencilerinin aşağıdaki şeklin alanını temsil etmek için birçok farklı ifadeyi gördüğünde şaşırıldı. Zeynep öğretmen gerçekte doğru olanları yanlış olarak işaretlediğinden emin olmak istedi. Aşağıdaki her bir ifade için verilen şeklin alanının doğru temsil edilip edilmediğine karar verin. (Her bir seçenek için TEMSİL EDER, TEMSİL ETMEZ veya EMİN DEĞİLİM şıklarından birini işaretleyin.)</p>																																																								
																																																									
<table border="1"> <thead> <tr> <th></th> <th>Correctly represents</th> <th>Does not correctly represent</th> <th>I'm not sure</th> </tr> </thead> <tbody> <tr> <td>a) a^2+5</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>b) $(a+5)^2$</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>c) a^2+5a</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>d) $(a+5)a$</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>e) $2a+5$</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>f) $4a+10$</td> <td>1</td> <td>2</td> <td>3</td> </tr> </tbody> </table>		Correctly represents	Does not correctly represent	I'm not sure	a) a^2+5	1	2	3	b) $(a+5)^2$	1	2	3	c) a^2+5a	1	2	3	d) $(a+5)a$	1	2	3	e) $2a+5$	1	2	3	f) $4a+10$	1	2	3	<table border="1"> <thead> <tr> <th></th> <th>Temsil Eder</th> <th>Temsil Etmez</th> <th>Emin Değilim</th> </tr> </thead> <tbody> <tr> <td>a) a^2+5</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>b) $(a+5)^2$</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>c) a^2+5a</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>d) $(a+5)a$</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>e) $2a+5$</td> <td>1</td> <td>2</td> <td>3</td> </tr> <tr> <td>f) $4a+10$</td> <td>1</td> <td>2</td> <td>3</td> </tr> </tbody> </table>		Temsil Eder	Temsil Etmez	Emin Değilim	a) a^2+5	1	2	3	b) $(a+5)^2$	1	2	3	c) a^2+5a	1	2	3	d) $(a+5)a$	1	2	3	e) $2a+5$	1	2	3	f) $4a+10$	1	2	3
	Correctly represents	Does not correctly represent	I'm not sure																																																						
a) a^2+5	1	2	3																																																						
b) $(a+5)^2$	1	2	3																																																						
c) a^2+5a	1	2	3																																																						
d) $(a+5)a$	1	2	3																																																						
e) $2a+5$	1	2	3																																																						
f) $4a+10$	1	2	3																																																						
	Temsil Eder	Temsil Etmez	Emin Değilim																																																						
a) a^2+5	1	2	3																																																						
b) $(a+5)^2$	1	2	3																																																						
c) a^2+5a	1	2	3																																																						
d) $(a+5)a$	1	2	3																																																						
e) $2a+5$	1	2	3																																																						
f) $4a+10$	1	2	3																																																						

The sample statements regarding the changes in the general cultural context and the number of changes are given in [Table 2](#) from the interviews with a group of experts in the field, which helps to determine whether the items reflect the situations that will arise in the classrooms in Türkiye after the item selection is translated.

Table 2. Exemplars of general contextual changes to items and frequency of changes.

Type of change	Example from original U.S. form	Example from adapted Turkish form	Number of items changed	
			Form A	Form B
People's names	Ms. Ashton	Asya öğretmen	18	20
	Ms. Diaz	Deniz öğretmen		
	Leah	Leyla		
	Earl	Enes		
Non-mathematical language	Baseball cards	Oyuncu kartı	5	6
	Mix contains	Çerez		

In [Table 2](#), it is seen that the changes in the context of general culture are evaluated in the sub-themes of "people's names and non-mathematical language". 18 changes were made in Form A, and 20 changes were made in Form B, which was developed in the context of people's names. There were 5 changes in form A and 6 changes in form B, which was developed in the context of non-mathematical language. Therefore, in this context, a total of 49 changes were made in the context of general culture, 23 changes in A form and 26 changes in B form.

The second stage in the cultural adaptation process includes changes in the context of school culture. Sample statements regarding the changes in this framework and the number of changes are presented in [Table 3](#).

Table 3. Examples of school contextual changes and frequency of changes in items

Changes' type	Original form	Adapted Turkish form	frequency of items changed	
			Form A	Form B
School language	Textbook	Ders kitabı	6	8
	Brainstorm	Beyin fırtınası		
	Ms. Hamid's class	Hatice öğretmenin öğrencileri		
Structure of education system	Kyle's method	Zeki'nin çözüm yolu	3	3
	Byron's approach to the problem	Burak'ın problem yaklaşımı		
	Task	Görev		

In [Table 3](#), it is seen that the changes in the context of school culture are evaluated in the sub-themes of "school language and structures of the education system". There were 6 changes in form A and 8 changes in form B, which was developed in the context of school language. In the context structures of the education system, 3 changes were made in the A form and 3 changes in the B form. Therefore, a total of 20 changes were made in the context of school culture, including 9 changes in Form A and 11 changes in Form B.

The third type of change in the adaptation process involves changes in the mathematical structure. Since the changes in this category do not disturb the mathematical structure of the items, the probability of affecting the mathematical knowledge of the test takers is very low. According to this, sample statements about the changes and the number of changes are presented in [Table 4](#).

Table 4. Exemplars of mathematical changes to items and frequency of changes.

Type of change	Example from original	Example from adapted	Number of items changed	
	U.S. form	Turkish form		
Symbolic notations	50 percent	%50	3	1
	n th	n.		
Mathematical language	Hexagon	Altıgen	29	36
	Doubling its length	Boyunun 2 katı		
Units of measurement	Area	Alan	5	7
	A half	Yarım		
	1 ounce	10 gr		

In Table 4, it is seen that the changes in the mathematical structure are evaluated in the sub-themes of "symbolic notations, mathematical language, and unit of measurement". 3 changes were made to the questions in the A form, and 1 change in the B form, which was developed in the context of the symbolic notations. There were 29 changes in the questions in the A form developed in the context of mathematical language and 36 changes in the B form. 5 changes were made to the questions in the A form developed in the context unit of measurement, and 7 changes were made to the B form. Therefore, a total of 74 changes were made in the context of the mathematical structure, 37 changes in the A form and 47 changes in the B form developed within this scope. After the changes in the measurement units of the developed test, the measurement units were converted to the metric units used in Türkiye. Thus, the adapted test has been changed to suit the mathematics culture of Türkiye as a result of the changes in the A Form and the B Form.

The fourth change in the adaptation process includes changes in the language structure. According to this, sample statements about the changes and the number of changes are presented in Table 5.

Table 5. Exemplars of language structure changes to items and frequency of changes.

Type of change	Example from original U.S. form	Example from adapted Turkish form	Number of items changed	
			Form A	Form B
Language structure	For each item	Her bir madde için	2	2
	Circle ONE answer	Sadece bir seçeneği işaretleyiniz		

It is essential to ensure the intelligibility of the items due to the changes that may occur in the language structure during the translation of the sentences or words in the test items into a different culture. The expression "For each item" in the original test is translated into Turkish as "her bir madde için". However, since the options in a multiple-choice test are expressed as "şık" in Turkish, the sentence is arranged as "her bir şık için".

3.2. Validity and Reliability

In order to examine the validity and reliability of the adapted version of a test, point biserial correlations were compared within the framework of psychometric analysis, item difficulty values, and reliability values of the forms were calculated.

3.2.1. Point-biserial correlation results (r_{pbi})

In Classical Test Theory, r_{pbi} was used to differentiate an item between respondents with higher mathematics teaching knowledge and those with low mathematics teaching knowledge. r_{pbi} for each item of the measurements of the A form of the PFA test in the Turkish context compared with the sample from the US teachers is given in Table 6.

Table 6. r_{pbi} for patterns MKT-PFA test (Form A and Form B) items ordered by estimates on Turkish Algebraic Teaching for Knowledge test (Form A and Form B).

Turkish r_{pbi} (Form A)	U.S. r_{pbi} (Form A)	Turkish r_{pbi} (Form B)	U.S. r_{pbi} (Form B)
0.321	0.440	0.185	0.420
0.512	0.543	0.288	0.523
0.425	0.443	0.192	0.231
0.451	0.493	0.345	0.350
0.504	0.753	0.491	0.491
0.164	0.220	0.410	0.478
0.337	0.632	0.313	0.534
0.345	0.442	0.405	0.540
0.441	0.745	0.336	0.346
0.444	0.341	0.368	0.423
0.375	0.598	0.336	0.560
0.258	0.285	0.431	0.567
0.301	0.333	0.426	0.506
0.423	0.498	0.500	0.625
0.416	0.499	0.381	0.602
0.450	0.696	0.479	0.747
0.384	0.575	0.492	0.762
0.264	0.489	0.414	0.755
0.404	0.433	0.368	0.848
0.402	0.775	0.275	0.286
0.358	0.694	0.434	0.634
0.394	0.700	0.513	0.659
0.321	0.670	0.453	0.595
0.362	0.428	0.323	0.328
0.418	0.554	0.356	0.709
0.414	0.595	0.242	0.379
0.338	0.513	0.427	0.588
0.134	0.131		
0.343	0.342		

For each item in form A of the PFA test, r_{pbi} was compared with the measurements in the Turkish and US contexts. According to Hambleton et al. (1991) criteria the correlation between r_{pbi} is high ($r=0.635$; $t=4,275$; $p<0.001$). With this result, the test can measure the intended characteristics of teachers and preservice teachers, as there is a high correlation between the measurements in the USA and the measurements in Türkiye. The following additions were made to the Results section: The correlation values of the 17th and 26th items in the adapted A form and the correlation values of the 1st, 3rd, 21st, and 27th items in the B form were found to be $< .3$. It is also noted that the correlation values obtained for the 26th item in the A form and the 3rd and 21st items in the B form, when compared to the US version, were also $< .3$.

For each item in form B of the PFA test, r_{pbi} was compared with the measurements in the Turkish and US contexts. According to Hambleton et al. (1991) criteria the correlation value between r_{pbi} is high ($r=0.6381$; $t=4.1438$; $p<0.001$). With this result, we can say that the test can measure the intended characteristics of teachers and preservice teachers, as there is a high correlation between the measurements in the USA and in Türkiye.

3.2.2. One-parameter IRT results

In the study, item difficulty and item discrimination values for each item in the forms are given in Table 8 by using a one-parameter IRT model to obtain the item difficulty of the items in the A and B forms of the MKT-PFA test.

Table 7. Item difficulties and discriminations of the items in the A and B forms of the MKT-PFA test.

Form A			Form B		
Item	Item difficulty	Item discrimination	Item	Item difficulty	Item discrimination
3	-2.474	.913	1	-2.603	.903
4a	-1.033	.782	2	2.970	.927
4b	-0.767	.676	6a	-0.954	.842
4c	-1.326	.816	6b	-2.016	.930
4d	-1.888	.927	6c	-2.064	.933
7	0.792	.636	6d	-1.006	.869
8	-2.760	.959	8	-3.713	.960
9a	-0.413	.572	13a	-0.588	.796
9b	-2.034	.945	13b	0.851	.830
9c	-0.578	.728	13c	0.603	.903
9d	-2.054	.937	15	0.954	.842
11	4.451	.409	16a	2.063	.854
14a	2.222	.461	16b	1.565	.793
14b	1.788	.779	16c	2.461	.892
14c	1.085	.682	16d	3.325	.945
14d	-2.762	.966	19a	-1.851	.830
15	2.328	.837	19b	-3.641	.958
17	3.102	.899	19c	-3.325	.945
18	1.612	.757	19d	-3.139	.972
20a	5.133	.974	21	1.150	.272
20b	6.042	.986	25a	4.242	.974
20c	6.639	.991	25b	4.476	.979
20d	5.813	.984	25c	4.043	.970
22a	-4.126	.948	25d	1.288	.751
22b	-2.558	.858	25e	3.325	.945
22c	-1.084	.828	27	1.071	.715
22d	-5.042	.985	28	-2.132	.908
26	-0.277	.549			
27	1.261	.708			

It is seen that the item discrimination indexes of the items are greater than .40. According to Brennan and NCME (2006), it can be said that the discrimination of all items in forms A and B is good. Item difficulty parameters reflect the differentiation states of the participants in the item process (Baker, 2001; de Ayala, 2013). For this reason, the labels used to define the discrimination of the substances in the MKT-PFA test can be associated with the value ranges of the parameters, as indicated in Table 8:

Table 8. Item difficulty distribution of the items in the A and B forms of the MKT-PFA test.

Level of difficulty	Form A	Form B
Very easy ($-4 \geq x$)	4	4
Easy ($-2 \geq x > -4$)	7	5
Moderate ($2 \geq x \geq -2$)	7	7
Hard ($4 > x > 2$)	6	6
Very hard ($x \geq 4$)	5	5

When the item difficulties are examined, it is seen that the items in the A and B forms show a normal distribution. The forms adapted to this distribution can distinguish those with high mathematical knowledge in the sample from those with low mathematical knowledge.

3.2.3. Reliability results

The reliability of the A form and B forms of the MKT-PFA test, that is, the KR-20 values of how consistent the scores of the respondents are over more than one item or multiple tests, are given in Table 9.

Table 9. Reliability of MKT-PFA test.

Form	Number of the items	KR-20 values	Lord reliability
PFA-A form	N=29	.712	.733
PFA-B form	N=27	.735	.756

The KR-20 value of the A form of the MKT-PFA test data obtained as a result of the application to the preservice teachers was calculated as .712, and the KR-20 value of the B form as .735. The Lord reliability of the A form of the MKT-PFA test of the data obtained as a result of the application to the preservice teachers was calculated as .733, and the Lord reliability of the B form as .756. The measurements obtained from the test are reliable with this value obtained. When Table 9 is examined, there is a difference between the reliabilities of Form A and Form B. The reason for this is that the number of items in Form A is more than the number of items in Form B.

3.2.4. Test Information Curve of A and B Forms of MKT-PFA Test

The test information curve expresses the level of knowledge at which the achievement test best measures individuals. Figure 3 shows the test data curves for Form A and Form B. The x-axis in the graphs is the scale score of the preservice teachers; 0 generally corresponds to the average preservice teacher in the population studied; Negative scores indicate less knowledgeable preservice teachers, and positive scores indicate more knowledgeable preservice teachers.

Figure 3. Test information curve of A and B forms of MKT-PFA test.

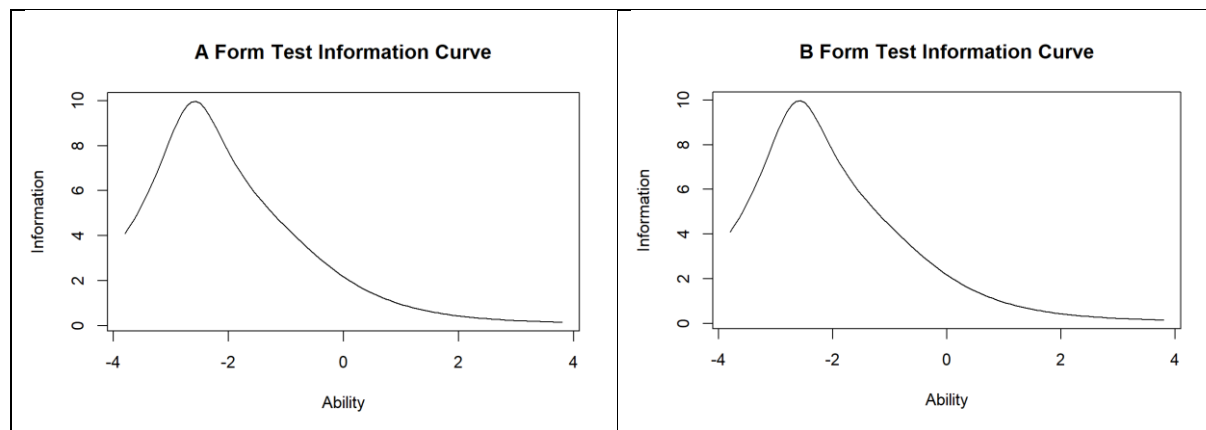


Figure 3 presents that the adapted form A and form B provide less information for preservice teachers who are 2.5 standard deviations above the mean and 1.5 standard deviations below the mean. Therefore, it means that form A and form B of the adapted test better distinguish preservice teachers with higher algebra knowledge for teaching from average or less algebra knowledge for teaching.

4. DISCUSSION and CONCLUSION

This section discusses the results of the analysis with qualitative and quantitative approaches in the adaptation process of the MKT-PFA test in the study. In this context, the cultural adaptation in the qualitative approach of the adaptation of the MKT-PFA test was analyzed in 4 categories.

The findings are discussed in each category. In the quantitative approach, the results obtained within the scope of the r_{pbi} , evaluating the item difficulties and reliability are discussed. Within the scope of this study, the Turkish adaptation of the items in the A and B forms of the MKT-PFA test was done by Delaney et al. (2008) following the steps given. Based on this research and similar studies, it can be said that the process of adapting such tests to a culture different from the one developed is arduous (Esendemir & Bindak, 2019; Marcinek et al., 2022; Ng, 2012; Ng et al., 2012).

4.1. Discussion of Studies in the Context of Cultural Adaptation of Test Items

The cultural adaptation process of the test items was carried out in four stages.

Changes in the context of general culture: The subject of general culture includes changing the non-mathematical but using daily language names and words in the test in a way that is suitable for the culture in question (Delaney et al., 2008). The use of food names in the question roots or options of the adapted test in mathematical problems serves to adapt to the cultural context. However, when adapting to a different country, the names of dishes or games in context may mean something different for the participants to whom the test will be applied (Ng, 2012). For this reason, cultural changes were made in this context in the study. While making changes, the names of similar foods were found without changing the mathematical situation in the problem. In addition, it was seen that changing the expression "baseball cards" to "playing cards" in the adapted culture in the adapted test did not make a semantic difference. For this reason, it did not create a change in the mathematical situation. Similarly, Ng (2012) adapted the word "pie" to their own culture as a cake or cake in their study.

Changes in the context of school culture: While adapting the items in the test, it was seen that the way of addressing the teachers differed between cultures. In the Turkish context, the term "teacher" is added next to the teacher's name, while in the Norwegian context, teachers are generally addressed by their first names. In addition, the expression of students in classes or groups is another matter of difference for different countries. For example, in Norway, there is an official statement that "classroom" should not be used when referring to student groups (Mosvold & Fauskanger, 2009). In addition to such differences in the context of school culture, it is seen that there are significant differences in teaching practices between cultures (Stiegler & Hiebert, 1999). In particular, the use of tangible materials as tools or models for representing mathematical ideas is different from the US in that many developing country teaching environments may not include physical manipulatives. For any of these manipulatives, for example, "Pattern Blocks" in the current study are clarified by providing either a description or a picture or both. While the context of the school culture is an important factor in determining the mathematical knowledge of the instructors, other factors also affect the mathematical knowledge required by the instructors.

Changes in the context of mathematical structure: Changes in the context of the mathematical structure of the items in the adapted test have the potential to lead to changes in the difficulty of the test (Delaney et al., 2008). Most symbolic expressions used in mathematics are universally acceptable. However, there may be differences between cultures regarding definitions or terminologies. Although technical terms such as "domino stones" or "mosaic" in the mathematical language context of the MKT-PFA test are available in Türkiye and the USA, these terms are not used at the primary level. Instead, a more general term, such as "pattern", is used. Such changes do not affect the integrity of the test in measuring their mathematical knowledge, as they are not the terms that teachers use in their teaching. For this reason, terms that measure teachers' familiarity with certain technical words may be preferred instead.

Similarly, Ng (2012) changed the term "polygon" to a more familiar term for Indonesian teachers, thus replacing it with "bangun datar segibanyak" meaning "multilateral flat shape". For Indonesian tutorials, this is a more descriptive term. Therefore, test items can be more understandable and easily adapted when evaluating teachers' knowledge of polygon definition.

While these differences in mathematical language do not pose a problem in the Irish context (Delaney et al., 2008), there are substantial variations within the context of mathematical language in test items in the examples of Türkiye, Korea, Indonesia, and Norway. There are also changes in the units of measurement. For example, while the unit of weight in the adapted A form was “ounces” in the developed context, it was changed to “package number” in the context of Türkiye.

Correspondingly, in the context of Indonesia and Norway, they expressed the measure of butter in the MKT-G test as “sticks” or “number of cups”. The situation for items, including money, is as follows: the fact that the difference between the currency in Türkiye and US has created a problem in terms of mathematical situations in the context of the items. For this reason, using equivalent values of money does not make mathematical sense. These two contexts are nearly impossible to translate into any of these languages without changing the entire context. This incomparable context problem poses a serious threat to the equivalence of the adapted data collection tool. Delaney et al. (2008) stated that there are relative similarities between the Irish and US forms of MKT forms, but there are differences in mathematical language, representation of concepts, measurement units, content and student knowledge. Although such differences can be ignored as they are mathematically insignificant, they indicate that differences in teachers’ performance on some items are sensitive to seemingly minor changes in items. For this reason, Delaney et al. (2008), Ng (2012), and Marcinek et al. (2022) stated that many changes can be made in the items of tests adapted to different cultures since the methods of teaching mathematics in cultures with different languages are significantly different.

Changes in language structure context: Delaney et al. (2008) stated minor language problems in the process of adapting the MKT test to Norway. They stated that these changes would not change the validity of the test items. However, additional explanations should be created to avoid confusion that may make the explanations at the root of the question or item in the test long and complex. As a result, all these factors should be considered when determining the mathematical knowledge of teachers and preservice teachers in different countries (Delaney et al., 2008).

4.2. Discussion of Findings Obtained from Psychometric Tests

After the cultural adaptation process for the items in the A and B forms of the MKT-PFA test was completed, point biserial correlation was obtained for each item in the A and B forms. It was concluded that r_{pbi} of the data was highly correlated between Türkiye and the United States. It was observed that there was a high level of correlation between the test adapted to Turkish and the test developed in the USA. It is seen that the correlations of some questions in the test adapted to Turkish are $<.3$. This situation also appears to be the case in the original form of MKT-PFA. In addition, these correlations in the study may be higher when working with larger sample groups. In addition, when we determine the subgroups in the low-correlation questions in the adapted form A and form B and look at the relationship at the class level, it is seen that the correlation value among the 4th grade preservice teachers is $>.3$. It was observed that there was a high correlation between the test adapted to Turkish and the test developed in the USA. In addition, using a one-parameter IRT model, it was seen that the distribution of the item difficulty values obtained for each item in the forms could distinguish between those with high mathematical knowledge in the sample and those with low mathematical knowledge in the adapted forms. Additionally, when the discrimination values of the items were examined, it was concluded that they were $>.40$. It can be said that form A and form B of the adapted tests can distinguish between preservice teachers who have good algebra knowledge for teaching and preservice teachers whose algebra knowledge for teaching is average or less. The test is reliable with these values obtained according to the KR-20 and Lord reliability values obtained for both forms of the MKT-PFA test. Finally, when the A form and B form Test Information Curve of the adapted test are examined, it is seen that the form better distinguishes the preservice teachers with higher algebra teaching knowledge from the preservice teachers with average or less

algebra teaching knowledge. In the pilot study for US teachers who participated in California's Mathematical Professional Development Institute, the MKT-PFA test Form A, Form B, and Form C provided maximum information for less knowledgeable teachers whose abilities are one-half standard deviation below the mean (Hill, 2007). The adapted Form A and Form B provide less information for preservice teachers who are 2.5 standard deviations above the mean and 1.5 standard deviations below the mean. Therefore, it means that Form A and Form B of the adapted test better distinguish preservice teachers with higher algebra knowledge for teaching from average or less algebra knowledge for teaching.

In this study, the MKT-PFA test was adapted to examine the mathematical knowledge of teachers and preservice teachers in Türkiye. In the adaptation process, the results of adaptation were included when translating test items from one language to another and for use in a different environment than intended. Delaney et al. (2008) suggest that international comparisons of teachers' mathematical knowledge should be evaluated in light of the differences that may exist in teachers' mathematical knowledge used in each country. Thus, clear guidelines should be developed to adapt the mathematical teaching information items. In addition, the differences in the mathematical knowledge of teachers or preservice teachers between countries can be explained by the differences in the mathematical knowledge used by teachers or preservice teachers in the relevant countries (Mosvold & Fauskanger, 2009; Ng, 2012). When comparing the knowledge of teachers between countries, it is insufficient to adapt the items from one country alone. For this reason, our study will shed light on the studies comparing different cultures with the Turkish context for the MKT-PFA test. Such research may lead to further development of the theoretical structure of MKT and possible cultural differences related to this structure. Additionally, Algebraic Knowledge for Teaching focuses on the knowledge and skills required for teachers or preservice teachers to improve their ability to explain and teach algebraic concepts to students. Ball et al (2008) discuss this special knowledge that teachers should have and how they can guide students' understanding of algebraic concepts. It aims to provide information to mathematics educators about the algebra teaching knowledge of preservice teachers with the adapted algebra knowledge for teaching tests. In this way, they can build "Algebra Teaching" courses aimed at the algebra teaching knowledge levels of preservice teachers. In addition, different instructional designs can be applied to better understand the relationship and interaction between mathematics teaching and MKT.

When we look at the results of the psychometric tests, the adaptation of the A and B forms of the MKT-PFA test is generally appropriate based on the psychometric analyses. In other words, a test developed to measure the mathematics knowledge of secondary mathematics teachers working in schools in the U.S. was successfully adapted to the Turkish context.

Acknowledgments

These results are a part of the University of Gaziantep doctoral thesis research project numbered EF.DT.22.04.

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, 21.03.2022-162213.

Contribution of Authors

Ali Bozkurt: Investigation, adaptation of the instruments and receiving experts' opinions.
Begüm Özmuşul: Adaptation of the instruments, data analysis, resources and visualization.

Orcid

Ali Bozkurt  <https://orcid.org/0000-0002-0176-4497>

Begüm Özmuşul  <https://orcid.org/0000-0003-0163-5406>

REFERENCES

- An, S., Kulm, G., & Wu, Z. (2004). The pedagogical content knowledge of middle school, mathematics teachers in China and the U.S. *Journal of Mathematics Teacher Education*, 7(2), 145-172.
- Aryadoust, V., Ng, L.Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, 38(1), 6-40. <https://doi.org/10.1177/0265532220927487>
- Baker, F.B. (2001). *The basics of item response theory* (2nd ed). (ED458219). <https://eric.ed.gov/?id=ED458219>
- Baker, F.B., & Kim, S.H. (2017). *The basics of item response theory using R* (Vol. 969). Springer.
- Ball, D.L. (1990). Prospective elementary and secondary teachers' understanding of division. *Journal for Research in Mathematics Education*, 21(2), 132-144.
- Ball, D.L., & Hill, H.C. (2008). Measuring teacher quality in practice. In D. H. Gitomer (Ed.), *Measurement Issues and Assessment for Teaching Quality*, pp. 80-98. SAGE.
- Ball, D.L., Hill, H.C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide? *American Educator* (Fall 2005), 14-46.
- 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>
- Brennan, R.L., & National Council on Measurement in Education (NCME). (2006). *Educational measurement*. Praeger.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University.
- Boykin, A.A., Ezike, N.C., & Mysers, A.J. (2023). Model-data fit evaluation: Posterior checks and Bayesian model selection. *International Encyclopedia of Education* (4th Edition), 279-289.
- Burnham, K.P., & Anderson, D.R. (2002). *Model selection and multimodal inference: a practical information-theoretic approach*. Springer.
- Charalambous, C.Y. (2008). *Prospective teachers' mathematical knowledge for teaching and their performance in selected teaching practices: Exploring a complex relationship*. (Doctoral dissertation) University of Michigan.
- Chou, Y.T., & Wang, W.C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement*, 70(5), 717-731. <https://doi.org/10.1177/0013164410379322>
- Cohen, I. (2011). Teacher-student interaction in classrooms of students with specific learning disabilities learning English as a foreign language. *Journal of Interactional research in communication disorders*, 2(2), 271-292. <https://doi.org/10.1558/jircd.v2i2.271>
- Council of Higher Education, (CoHE (Yükseköğretim Kurulu), 2018). *New teacher training programs, reasons for updating the programs, innovations and implementation principles* [In Turkish]. https://www.yok.gov.tr/Documents/Kurumsal/egitim_ogretim_dairesi/Yeni-Ogretmen-Yetistirme-Lisans_Programlari/AA_Sunus_%20Onsoz_Uygulama_Yonergesi.pdf [In Turkish]
- Cole, Y. (2012). Assessing elemental validity: The transfer and use of mathematical knowledge for teaching measures in Ghana. *ZDM*, 44(3), 415-426. <https://doi.org/10.1007/s11858-012-0380-7>
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. (ED312281). ERIC. <https://eric.ed.gov/?id=ED312281>
- Cronbach, L.J., & Shavelson, R.J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-418.

- Çelen, Ü. (2008). Comparison of validity and reliability of two tests developed by classical test theory and item response theory. *Elementary Education Online*, 7(3), 758-768.
- de Ayala, R.J. (2013). *The theory and practice of item response theory*. Guilford.
- Delaney, S., Ball, D.L., Hill, H.C., Schilling, S.G., & Zopf, D. (2008). Mathematical knowledge for teaching: Adapting US measures for use in Ireland. *Journal of Mathematics Teacher Education*, 11(3), 171-197. <https://doi.org/10.1007/s10857-008-9072-1>
- Driscoll, M. (1999). *Fostering algebraic thinking: a guide for teachers grades 6-10*. NH: Heinemann.
- Edelen, M.O., & Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of life research*, 16, 5-18. <https://doi.org/10.1007/s11136-007-9198-0>
- Embretson, S.E., & Reise, S.P. (2013). *Item response theory*. Psychology.
- Esendemir, O., & Bindak, R. (2019). Adaptation of the test developed to measure mathematical knowledge of teaching geometry in Turkey. *International Journal of Educational Methodology*, 5(4), 547-565. <https://doi.org/10.12973/ijem.5.4.547>
- Fan, J., & Bond T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment, Vol. I: Fundamental techniques* (pp. 83–102). Routledge. <https://doi.org/10.4324/9781315187815>
- Fauskanger, J., Jakobsen, A., Mosvold, R., & Bjuland, R. (2012). Analysis of psychometric properties as part of an iterative adaptation process of MKT items for use in other countries. *ZDM*, 44, 387–399. <https://doi.org/10.1007/s11858-012-0403-4>
- Frary, R.B. (1989). Partial credit scoring methods for multiple choice Tests. *Applied Measurement in Education*, 2(1), 79-96.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. SAGE Publications.
- Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Academic Publishers Group.
- Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–240.
- Han, H. (2022). The effectiveness of weighted least squares means and variance adjusted based fit indices in assessing local dependence of the rasch model: Comparison with principal component analysis of residuals. *PloS ONE*, 17(9). <https://doi.org/10.1371/journal.pone.0271992>
- Hill, H.C. (2007). Mathematical knowledge of middle school teachers: Implications for the no child left behind policy initiative. *Educational Evaluation and Policy Analysis*, 29(2), 95–114. <https://doi.org/10.3102/0162373707301711>
- Hill, H.C., & Ball, D.L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education*, 35(5), 330-351.
- Hill, H.C., Rowan, B., & Ball, D.L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research journal*, 42(2), 371-406. <https://doi.org/3699380>
- Hill, H.C., Schilling, S.G., & Ball, D.L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal*, 105(1), 11-30. <https://doi.org/10.1086/428763>
- Hill, H., & Ball, D.L. (2009). The curious and crucial case of mathematical knowledge for teaching. *Phi Delta Kappan*, 91(2), 68-71. <https://doi.org/10.1177/00317217090910021>
- Holmes, F. & Brian F.F. (2019). A comparison of estimation techniques for IRT models with small samples, *Applied Measurement in Education*, 32(2), 77-96. <https://doi.org/10.1080/08957347.2019.1577243>

- Huang, R., & Kulm, G. (2012). Preservice middle grade mathematics teachers' knowledge of algebra for teaching. *The Journal of Mathematical Behavior*, 31(4), 417-430. <https://doi.org/10.1016/j.jmathb.2012.06.001>
- Kieran, C., Kieran, C., & Ohmer. (2018). *Teaching and learning algebraic thinking with 5-to 12-year-olds* (pp. 79-105). Springer.
- Kim, Y. (2016). Interview prompts to uncover mathematical knowledge for teaching: focus on providing written feedback. *The Mathematics Enthusiast*, 13(1), 71-92. <https://doi.org/10.54870/1551-3440.1366>
- Kim, Y. (2020). Korean teachers' mathematical knowledge for teaching in algebraic reasoning. *Journal of Educational Research in Mathematics, (Special Issue)*, 185-198. <https://doi.org/10.29275/jerm.2020.08.sp.1.185>
- Knipping, C. (2003). Learning from comparing. *Zentralblatt für Didaktik der Mathematik*, 35(6), 282-293.
- Kline, P. (1994) *An Easy Guide to Factor Analysis*. Routledge.
- Koellner, K., Jacobs, J., Borke, H., Schneider, C., Pittman, M.E., Eiteljorg, E., & Frykholm, J. (2007). The problem-solving cycle: A model to support the development of teachers' professional knowledge. *Mathematical Thinking and Learning*, 9(3), 273-303. <https://doi.org/10.1080/10986060701360944>
- Kwon, M., Thames, M.H., & Pang, J. (2012). To change or not to change: Adapting mathematical knowledge for teaching (MKT) measures for use in Korea. *ZDM*, 44, 371–385. <https://doi.org/10.1007/s11858-012-0397-y>
- Langrall, C.W. & Swafford J.O. (1997). Grade six students' use of equations to describe and represent problem situation. *Paper presented at the American Educational Research Association*, Chicago, IL.
- Lew, H.C. (2004). Developing algebraic thinking in early grades: Case study of Korean elementary school mathematics. *The Mathematics Educator*, 8(1), 88-106.
- Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Pub. Co.
- Ma, L. (1999). *Knowing and teaching elementary mathematics: Teachers' understanding of fundamental mathematics in China and the United States*. Lawrence Erlbaum Associates, Inc.
- Marcinek, T., & Partová, E. (2016). Exploring cultural aspects of knowledge for teaching through adaptation of U.S.-developed measures: Case of Slovakia. Paper presented at the *13th International Congress on Mathematical Education*. Hamburg, Germany.
- Marcinek, T., Jakobsen, A., & Partová, E. (2022). Using MKT measures for cross-national comparisons of teacher knowledge: case of Slovakia and Norway. *Journal of Mathematics Teacher Education*, 1-31. <https://doi.org/10.1007/s10857-021-09530-3>
- Morris, A.K., Hiebert, J., & Spitzer, S.M. (2009). Mathematical knowledge for teaching in planning and evaluating instruction: What can preservice teachers learn?. *Journal for research in mathematics education*, 40(5), 491-529. <https://doi.org/10.5953/jrme.40.5.491>
- Mosvold, R., & Fauskanger, J. (2009). *Challenges of translating and adapting the MKT measures for Norway*. Paper presented at the American Educational Research Annual Meeting in San Diego, CA.
- National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. U.S. Department of Education.
- Ng, D. (2012). Using the MKT measures to reveal Indonesian teachers' mathematical knowledge: Challenges and potentials. *ZDM*, 44(3), 401-413. <https://doi.org/10.1007/s11858-011-0375-9>
- Ng, D., Mosvold, R., & Fauskanger, J. (2012). Translating and adapting the mathematical knowledge for teaching (MKT) measures: The cases of Indonesia and Norway. *The Mathematics Enthusiast*, 9(1), 149-178. <https://doi.org/10.54870/1551-3440.1238>

- Özdemir, D. (2004). A comparison of psychometric characteristics of multiple choice tests based on the binaries and weighted scoring in respect to classical test and latent trait theory. *Hacettepe University Journal of Education*, 26, 117-123.
- Pekmezci, F.B., & Avşar, A.Ş. (2021). A guide for more accurate and precise estimations in Simulative Unidimensional IRT Models. *International Journal of Assessment Tools in Education*, 8(2), 423-453. <https://doi.org/10.21449/ijate.790289>
- Reyhanlıoğlu, Ç., & Doğan, N. (2020). An analysis of parameter invariance according to different sample sizes and dimensions in parametric and nonparametric item response theory. *Journal of Measurement and Evaluation in Education and Psychology*, 11(2), 98-112. <https://doi.org/10.21031/epod.584977>
- Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, 17(1), 321-33. <https://doi.org/10.12738/estp.2017.1.0270>
- Schmittau, J. (2005). The development of algebraic thinking. *Zentralblatt für Didaktik der Mathematik*, 37(1), 16-22.
- Sheng, Y. (2013). An empirical investigation of Bayesian hierarchical modeling with unidimensional IRT models. *Behaviormetrika*, 40(1), 19-40. <https://doi.org/10.2333/bhmk.40.19>
- Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4-14.
- Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of test let-based tests. *Journal of Educational Measurement*, 28(4), 237-247.
- Smith, R.M., & Miao, C.Y. (1994). Assessing unidimensionality for Rasch measurement. *Objective Measurement: Theory into Practice*, 2, 316-327.
- Stiegler, J.W., & Hiebert, J. (1999). *The Teaching Gap. Best ideas from the world's teachers for improving education in the classroom*. The Free.
- Strand, K., & Mills, B. (2014). Mathematical content knowledge for teaching elementary mathematics: A focus on algebra. *The Mathematics Enthusiast*, 11(2), 385-432. <https://doi.org/10.54870/1551-3440.1307>
- Tabachnick, B., & Fidell, L. (2012). *Using multivariate statistics*. Pearson.
- Welder, R.M., & Simonsen, L.M. (2011). Elementary Teachers' Mathematical Knowledge for Teaching Prerequisite Algebra Concepts. *Issues in the Undergraduate Mathematics Preparation of School Teachers*, 1.
- Wilson, L., Andrew, C., & Sourikova, S. (2001). Shape and structure in primary mathematics lessons: A comparative study in the North-east of England and St Petersburg, Russia-some implications for the daily mathematics lesson. *British Educational Research Journal*, 27(1), 29-58.
- Wright, B.D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10(3), 509–511. <https://www.rasch.org/rmt/rmt103b.htm>
- Yang, S. (2007). *A comparison of unidimensional and multidimensional RASCH models using parameter estimates and fit indices when assumption of unidimensionality is violated* [Doctoral dissertation]. The Ohio State University.
- Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.
- Zazkis, R., & Liljedahl, P. (2002). Generalization of patterns: The tension between algebraic thinking and algebraic notation. *Educational Studies in Mathematics*, 49(3), 379-402.

APPENDIX

Some released items from MKT - test

1. Zeliha ve öğretmeni Zeliha'nın doğum gününde beraber kurdukları aşağıdaki problemi sınıf arkadaşlarına sormuşlardır:

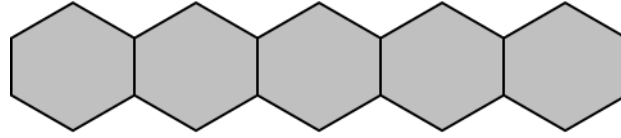
Zeliha'nın şimdiki yaşı erkek kardeşinin yaşının iki katıdır. Kaç yıl sonra Zeliha'nın yaşı kardeşinin yaşının yine iki katı olur?

Arkadaşları aşağıdaki cevapları vermiştir. Bu cevaplardan hangisini doğru olarak kabul edersiniz? (Sadece bir şıkkı işaretleyiniz.)

- A) Her 2 yılda bir olur.
- B) Zeliha'nın yaşına bağlıdır.
- C) Zeliha'nın yaşı, şimdiki yaşının 2 katı olduğunda olur.
- D) Bir daha asla olmaz.

2. Jale öğretmen dersinde kullanmak için şu problemi hazırlamıştır:

Aşağıdaki gibi bir satıra yan yana 100 düzgün altıgeni dizerseniz oluşan şeklin çevresi ne olur?



Jale öğretmen öğrencilerinden gelebilecek farklı çözümleri görmek istemiştir. Aşağıda verilmiş muhtemel öğrenci çözümlerinden hangileri doğru cevaba götürür? (Her bir şık için evet, hayır veya emin değilim seçeneklerinden birini işaretleyiniz.)

	Evet	Hayır	Emin Değilim
A) $4 \times 100 + 2$	1	2	3
B) $(6 \times 100) - 2 \times 99$	1	2	3
C) $4 \times 98 + 2 \times 5$	1	2	3
D) 6×100	1	2	3

The mental imagery scale for art students: Building and validating a short form

Handan Narin Kızıltan^{1*}, Hatice Cigdem Bulut²

¹Cukurova University, Faculty of Education, Department of Art Education, Adana Türkiye

²Northern Alberta Institute of Technology, Education Insights, Data & Research, Edmonton, AB Canada

ARTICLE HISTORY

Received: Apr. 30, 2024

Accepted: July 21, 2024

Keywords:

Mental imagery,

Scale shortening,

Ant colony optimization,

Item selection,

Psychometrics.

Abstract: Mental imagery is a vital cognitive skill that significantly influences how reality is perceived while creating art. Its multifaceted nature reveals various dimensions of creative expression, amplifying the inherent complexities of measuring it. This study aimed to shorten the Mental Imagery Scale in Artistic Creativity (MISAC) via the Ant Colony Optimization algorithm (ACO), a metaheuristic methodology for developing psychometrically robust brief scales. Answering 63 items in the original version of MISAC demands a higher cognitive load and, consequently, more time. Therefore, our goal was to shorten it while preserving its psychometric properties. In this study, responses to the MISAC were obtained from 500 undergraduate students enrolled in an art education program. The items on the short form of the MISAC were selected based on pre-specified validity criteria and content representability. The 28-item short form of MISAC demonstrated comparable performance to the original version regarding construct validity, criteria-related validity, and reliability coefficients. Moreover, strict invariance was attained across both gender groups in the validation process of the short form. These results highlight the utility of the shortened version of the MISAC as a valid measure with minimal loss of information of scores compared to the full version.

1. INTRODUCTION

Mental imagery, considered one of the critical cognitive skills for humans (Pérez-Fabello & Campos 2007), plays a crucial role in the perception of reality during the artistic production process (Ziss, 2011). Mental imagery occurs when perceptual information is accessed from memory and can be created by combining and manipulating stored perceptual information in new ways (Kosslyn et al., 2001). Therefore, mental images include both visual representations and various types of past mental encounters (Hilton, 2007).

Following the second half of the 20th century, interest in mental imagery has accelerated in fields such as behavioral and cognitive psychology, clinical psychology, neuroscience, marketing, and sport (Kosslyn et al., 2001; Park & Yoo, 2020; Pearson et al., 2015; Saulsman et al., 2019). In psychological research, mental imagery is utilized to prevent mental disorders and develop treatment methods (Saulsman et al., 2019; Schwarz et al., 2020). In addition, it has been used to

*CONTACT: Handan NARİN KIZILTAN ✉ handannarinn@gmail.com 📧 Cukurova University, Faculty of Education, Department of Art Education, Adana, Türkiye

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

change individuals' psychological attitudes, perceptions, and perspectives (Holmes & Mathew, 2010; Park & Yoo, 2020; Pearson, 2019; Saulsman et al., 2019). Mental imagery is also a crucial cognitive domain often highlighted in art education (Duncun, 2001; Heid et al., 2009) owing to its relation to creativity (Palmiero et al., 2016). Consequently, art and cognition are intricately linked, mutually reflecting and reinforcing each other (Bhattacharya & Petsche, 2002).

Artists create art by drawing on mental images developed through observing the world (Hetland et al., 2007). The personal records, diaries, and sketchbooks of well-known artists like Leonardo Da Vinci and Picasso, which reveal their internal worlds, demonstrate that they actively utilized imagery while creating their artwork. (Rosenberg, 1987; Vellera & Gavard-Perret, 2012). This same process applies to art students in visual arts classes. Art students learn to observe and use observation to generate mental images and plan ways to create their artwork (Hetland et al., 2007). Therefore, the power of imagination is an intrinsic and essential element for art students (Bhattacharya & Petsche, 2002; Chamberlain et al., 2019).

Imagination is a product of cognitive actions that facilitate the construction of new meanings (Efland, 2002). Metaphorical thinking is one of these cognitive actions utilized to imbue meaning in creating and evaluating artwork (Hetland et al., 2007; Serig, 2006). The tools artists employ during artwork production are integral to the cognitive process used by those interpreting the artwork to construct meanings (Efland, 2002). Metaphorical thinking involves expressing different concepts through a single, similar concept that can be represented in various ways (Deaver & Shiflett, 2011). It directs minds beyond existing similarities to new similarities it creates, leading to the discovery of a new dimension of meaning for the word (Lakoff & Johnson, 2010).

Consequently, previously undiscovered creative meanings are brought forth. When engaging in metaphorical thinking or drawing, an individual participates in the form of mental imagery (Dodson, 2013). Images also serve as metaphorical conceptualizations and a creative act of reinterpreting these concepts. Creative thinking involves the cognitive properties of metaphor capable of generating new meanings by establishing connections between different elements (Efland, 2002). Images are both a metaphorical conceptualization and a creative action. Therefore, when it comes to artistic creativity, a robust relationship exists between mental imagery, metaphorical thinking, and creative thinking. For this reason, the ability to form mental images can be associated with the data obtained from the metaphorical thinking ability test, which measures the ability to produce meaning, and the drawing test, which measures the ability to create creative images.

Given the ongoing importance and long history of mental imagery within art, there have been various studies focusing on mental imagery in art (Jankowska & Karwowski, 2015; Pérez-Fabello & Campos, 2007; Pérez-Fabello et al., 2016). Furthermore, the relationship between creativity in art and mental imagery is examined in several studies (Miller, 2014; Pérez-Fabello & Campos, 2007; Pérez-Fabello et al., 2016; Vellera & Gavard-Perret, 2012). A study by Drake et al. (2021) found that artists possess superior imagery skills compared to non-artists, as assessed by a self-report measure. Another study (Vellera & Gavard-Perret, 2012) found that an increase in mental imagery score corresponded with an increase in performance in creative tasks, as measured by two different tools. In another study by Jankowska and Karwowski (2020), the results from five separate studies, each employing various measurement tools, were combined. The study found that art students exhibited a higher level of mental imagery compared to the non-artist group. These studies provide evidence that mental imagery is considered an indicator of artistic creativity by using different measures.

There are primarily three ways to assess mental imagery (Ji et al., 2019): (a) Reporting naturally occurring mental imagery, (b) Laboratory assessments of mental imagery, and (c) Scales for mental imagery. Applying scales in the fields of art and creativity can be more convenient for researchers due to the focus in these fields not typically being placed on the neurocognitive basis of mental imagery. However, the construct of mental imagery has been a challenge to measure both validly and reliably due to its multidimensional nature (Cumming & Eaves, 2018). In the literature, a variety of measures are focused on different aspects of mental imagery (e.g., Betts' Questionnaire Upon Mental Imagery, [Betts' QMI; Betts, 1909]; Vividness of Visual Imagery Questionnaire

[VVIQ; Marks, 1973]; The Plymouth Sensory Imagery Questionnaire [Andrade et al., 2014]). However, the multidimensional structure of mental imagery requires the use of more than one measurement tool or longer measures, which include several factors (Calabrese & Marucci, 2006; Jankowska & Karwowski, 2015; Vellera & Gavard-Perret, 2012).

Given the drawbacks of longer measures, such as decreasing response rate and increasing response bias (e.g., careless responding [Niessen et al., 2016], exhibiting response styles [Weijters et al., 2010]), researchers conducting similar studies prefer using shorter measures or short versions of commonly utilized and adapted scales (e.g., short versions of Betts' QMI, Sheehan (1967), and VVIQ; Marks, 1995). As a result, scale-shortening procedures have recently gained popularity in psychological and cognitive assessments due to the development of automated methods (Basarkod et al., 2018; Schroeders et al., 2016).

This study aimed to shorten the Mental Imagery Scale in Artistic Creativity (MISAC), which was recently developed in art education. We employed methodological advances in scale-shortening techniques and utilized a metaheuristic approach (e.g., Ant Colony Optimization algorithm [ACO]) to shorten the MISAC. In addition, we gathered reliability and validity evidence for the shortened version of the MISAC. Also, we compared the psychometric features of the full version of the MISAC with that of the shortened version.

1.1. The MISAC

The MISAC measures the ability of individuals to recreate/remember objects, events, and phenomena based on their physical (movement, shape, color, place) and sensory modalities (e.g., sound, texture, and taste) (Narin, 2019). In this context, the scale measures the mental imagery ability of spatial, tactile, physical, kinesthetic, emotional, characteristic features, and affective experiences. While developing the MISAC, Mark's VVIQ (Mark, 1973) and Sheehan's Betts' QMI (1967) scales were considered. Mark's VVIQ scale includes four different contents (i.e., visualizing sentences about relatives or friends, the sunrise, a shop one often goes to, and the image of a country). Sheehan's Betts' QMI (1967) includes sensory modalities: visual, auditory, tactile (cutaneous), kinesthetic, gustatory, olfactory, and organic (whole body).

Unlike the scales mentioned above measuring the vividness of mental imagery, the MISAC measures the ability of mental imagery in terms of vividness, attention, and control. The most distinctive difference between the MISAC and other scales is its use in determining the mental imagery ability of a group within programs that require artistic creativity or in creative individuals such as those enrolled in art education programs. Notably, the MISAC can be utilized as a supplementary measurement tool for art and creativity research and for the selection procedures of students entering arts education or art-related programs. It can also be utilized to follow students' progress in different disciplines that require creative skills, such as visual communication design, art and design, and architecture.

Use of the MISAC not only considers the insights of mental imagery scales from working with participants with differing characteristics and creative individuals (Kozhevnikov et al., 2013; Miller, 2014; Pérez-Fabello et al., 2016; Vellera & Gavard-Peret, 2012) but also considers the limitations of current scales and attempts to overcome their shortcomings. For example, Sheehan's (1967) QMI contains smell as one of the sensory modalities; however, Arshamian and Larsson (2014) indicated that individuals, in most cases, cannot produce mental images based on the sense of smell. In addition, Kozhevnikov et al. (2013) noted that despite the importance of the ability to visualize and discriminate colors and textures of objects for artistic creativity, these aspects are often neglected in the current measures. Thus, the MISAC incorporates various conceptualizations regarding mental imagery within its list of items and factors.

Based on the original version of the MISAC, comprising seven factors and 63 items (Narin, 2019), the exploratory factor analysis (EFA) results revealed that the scale accounted for 49.6% of the total variance, with factor loadings values ranging from .45 to .74. The confirmatory

factor analysis (CFA) supported the factor structure of the MISAC, as evidenced by good fit values (RMSEA = .05, NFI = .90, NNFI = .95, CFI = .95, SRMR = .06, IFI = .95) (χ^2 (1869) = 3525.56, $p < .001$). In the original version of the MISAC, each factor exhibited good internal consistency, with Cronbach's alpha values ranging between .82 and .89 (Narin, 2019).

1.2. Why Shorten The MISAC?

Long measures may cause fatigue, higher drop-out rates, and a lower response rate, as well as increase an unnecessary waste of time and energy, thereby reducing the quality of the gathered data (Basarkod et al., 2018; Olaru et al., 2015; Rammstedt & Beierlein, 2014). Also, if longer measures include items demanding a higher cognitive load, as seen in MISAC, then the undesired effects may be problematic regarding data quality. Responding to items regarding mental imagery might take longer than responding to items in other settings, as one must imagine the vividness of the object being questioned within an item. When items get more cognitively demanding, the respondents may likely adapt their response style as a shortcut (Krosnick et al., 2002). Thus, the length of scales and the level of cognitive load may be obstacles to obtaining the intended data quality.

Studies using mental imagery scales aim to determine the associations with other variables (Jankowska & Karwowski, 2015; Pérez-Fabello & Campos, 2007; Pérez-Fabello et al., 2016). Therefore, respondents may be required to respond to several questionnaires to provide scholars with a wide array of information regarding their visual and mental abilities. Due to assessment time and research funding sometimes being limited in designs that include multiple constructs, keeping the response rate and costs at a reasonable level is important, so shorter scales are more preferable (Rammstedt & Beierlein, 2014; Dogan & Bulut, 2024). Therefore, developing shorter versions of some scales has steadily increased over the past few years to eliminate these consequences.

In numerous higher education institutions, including those in Türkiye (e.g., O'Donoghue, 2011; Ozmutlu & Tomak, 2021; Taskesen, 2019; Tay, 2019; Yilmaz, 2016), scales or tests assessing creativity or related constructs hold significance in the selection process for art students, often complementing the evaluation of portfolios. However, the inclusion of multiple assessments, particularly longer ones, poses challenges for both candidates and the academic jury overseeing the selection process. This extended evaluation complicates the assessment for candidates and creates difficulties for the jury in making decisions based on these assessments. In response to these challenges, Turkish institutions frequently depend on evaluating drawing skills, including drawings of live models and imaginative design studies. (e.g., Dilmac & Kucuoglu, 2010; Taskesen, 2019); however, this approach introduces its own validity concerns. Including longer assessments, especially those focused on visualization skills, adds complexity to achieving thorough, reliable, and valid evaluations. Finding instruments that balance brevity with comprehensive assessment and validity poses a significant challenge. A shorter MISAC version can be a potential solution to bridge this evaluative gap.

1.2.1. Ant-Colony optimization

There has been a growing interest in the methods of automated approaches to scale shortening (Leite et al., 2008; Olaru et al., 2015; Schroeders et al., 2016; Yarkoni, 2010). The traditional approaches consist of examination of item-total correlations (Bowns et al., 2022; Carr et al., 2005), conducting EFA and choosing the highest factor loadings (Botes et al., 2021; Leite et al., 2008) researchers select items based on a reduction in the scale's Cronbach's alpha reliability coefficient if each item is removed (e.g., Bowns et al., 2022; Swindle et al., 2006). Inevitably, selecting the appropriate items via traditional approaches can take some time. More importantly, sequence effects or relying solely on one criterion within the abbreviating process result in unwanted biases. Furthermore, the required input from researchers is relatively high compared to automated approaches, depending on the number of items and factors on instruments and the number of criteria researchers consider (Yarkoni, 2010). Therefore, the

workload can be a significant obstacle in this process. On the other hand, traditional scale shortening methods may miss the most optimal version, as researchers only consider limited alternate forms. As a result, not only can automation significantly reduce time spent on developing short measures, but also it allows researchers to achieve optimality or near-optimality (Jankowsky et al., 2020; Olaru et al., 2019; Raborn et al., 2019; Sandy et al., 2014).

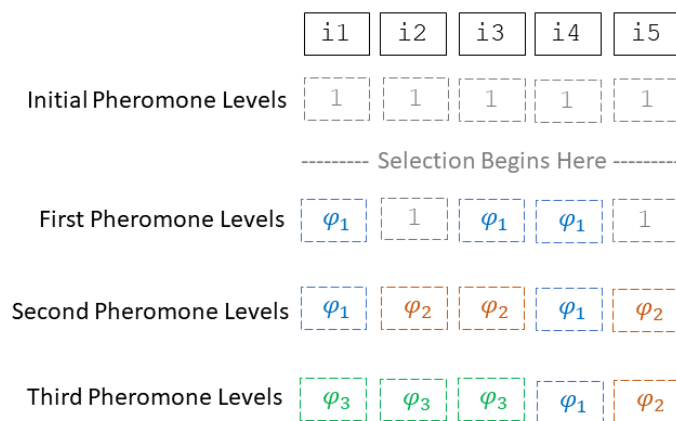
Automated approaches for scale abbreviation, such as Genetic Algorithms or Ant-Colony Optimization, make this process much faster and easier (Leite et al., 2008; Yarkoni, 2010). For instance, let us consider a situation where a researcher wants to shorten a 63-item with a seven-factor scale and use several criteria. If all 63 items of the long version are evenly distributed across the seven subscales (i.e., 9 items per subscale) and a short measure is constructed with 4 items per subscale, this would result in $\binom{9}{4}^7 = 504,189,521,813,376$ possible combinations.

After finding a suitable short form, the researcher may have to perform additional analysis for other criteria. However, using an automated approach, the process can be done more efficiently (Sandy et al., 2014; Yarkoni, 2010), and the researcher can use their time and expertise to evaluate the results instead of conducting multiple analyses. Research shows that automated approaches can provide better results than traditional approaches (Leite et al., 2008; Raborn et al., 2019; Sandy et al., 2014). For instance, Sandy et al. (2014) compared one rational approach and an automated approach (genetic algorithm approach) to develop a short scale. The validity and reliability properties of the scales developed separately by these approaches were similar. Similarly, Leite et al. (2008) showed that ACO excels at maximizing certain predefined qualities and outperforms methods for selecting items with traditional methods.

The current study used the Ant Colony Optimization (ACO) approach (Marcoulides & Drezner, 2003) to shorten the MISAC. The goal of using the ACO algorithm approach in this study was twofold. The first objective was to create a shortened version of MISAC to make it more practical for researchers to use in studies that do not have the capacity to use longer measures. Second, we wanted to maximize the model fit of a short form of MISAC in terms of converging on the previously validated mental imagery model. The ACO is one of the best-performing practices for producing short forms (Leite et al., 2008; Olaru et al., 2015; Raborn et al., 2019).

Interestingly, the ACO is a heuristic algorithm that incorporates the foraging behaviors of real ants to establish the shortest route to a food source in an automated model-fitting process (Marcoulides & Drezner, 2003). Deneubourg et al. (1983) found that ants produce pheromones while searching for a food source so that the ants that come after can utilize this chemical trail as feedback for determining the shortest path to the located food source. For example, ants will randomly try routes in the first step and produce pheromone chemicals during the search for routes to a food source. When a route is relatively long, its pheromone level will gradually dissipate, ultimately failing to attract other ants. Similarly, pheromone evaporation in the ACO algorithm can reduce the strength of pheromone routes over time. The evaporation rate can impact how well the ACO algorithm performs. This rate can encourage greater exploration of the solution space. However, it can also lead the algorithm to rapidly forget earlier successful solutions or prompt ants to follow existing routes more frequently, thus increasing the likelihood of the algorithm adhering to previously found shorter paths. At the end of this process, ants try to choose the shortest route over time.

In survey research, the ACO mimics those behaviors to generate short forms of scales by using the ‘pheromone’ levels of items (Olaru et al., 2015). For this, random models are generated through the ACO to determine the pheromone levels of items in the first iterations. Then, items that show the best fit in terms of specific criteria (i.e., model fit statistics) have higher probabilities of being selected in later iterations (Olaru et al., 2015). The process is complete once all the criteria are met by the number of items required for the short form. Figure 1 illustrates these steps in an example.

Figure 1. An illustration of the item selection procedure using the ACO.

In the initial stage, all items in this sample scale (i.e., i1, i2, i3, i4, and i5) have equal initial weights for the selection procedure. After selection begins in the ACO algorithm, a randomly short form is generated by selecting items 1, 3, and 4 and their pheromone levels (φ_1) are calculated for that short form. In the initial iteration, the ACO algorithm randomly selects items 1, 3, and 4, subsequently evaluating their suitability based on pheromone levels. These levels are calculated based on the criteria introduced to the algorithm (e.g., CFI > .95 and RMSEA < .06). These criteria can be various and are up to researchers and scale properties. Then, the pheromone levels influence and modify the weighting or significance of the selected items within the selection process (Leite et al., 2008). The algorithm integrates a pheromone evaporation mechanism, which reduces the current pheromone levels before adjusting them according to a pre-established rate determined by the researcher. This rate selection is pivotal, as it directs the algorithm's inclination towards favoring frequently selected items or encouraging greater exploration of potential item combinations in each iteration. Consequently, this step significantly contributes to fine-tuning the item selection process, emphasizing the influence or reliance on previously chosen items. In our five-item scenario, the process repeats for the second and third selections until the best items are chosen. If the third round marks the end, using the calculated pheromone levels helps identify the most suitable items based on how many the researcher aims to include in their shorter scale. As researchers can decide the criteria (e.g., model fit, number of items for each factor) and the parameters (i.e., number of ants and evaporation rate) to be introduced in the algorithm, the ACO provides flexibility and rapid solutions for the scale abbreviation process.

2. METHOD

2.1. Sample

The sample participants comprised 500 undergraduate students (29.2% males) aged 18-47 ($M=22.3$, $SD = 3.86$). The study recruited participants from five higher education institutions located in three different cities in Turkey, all of which specialize in providing education in the arts. The participants were drawn from the Fine Arts Education Department of the Education Faculty, as well as the Painting, Graphics, and Sculpture Departments of the Fine Arts Faculty. These departments were selected because they highly emphasize creativity and visual skills, which are essential for success and acceptance in the field. The number and percentage of first-year students, sophomores, juniors, and seniors were 119, 124, 125, 132 (23.8%, 24.8%, 25.0%, 26.4%), respectively. Before participating in the study, each student was given a detailed description of the research and asked to provide informed consent.

In this study, we present the results about the full scale and its properties ($N = 420$), obtained from a separate study (Narin, 2019). This sample shares resemblances with the sample characteristics employed in the current study. These undergraduates belong to the same

programs, encompassing approximately 28% male students, with an approximate 27% distribution across each academic year, ranging from first-year students to senior student cohorts.

2.2. Instruments

The MISAC, consisting of 63 items, is utilized to assess the mental imagery of art education students and help to evaluate how clearly and vividly people remember various objects, situations, facts, and events, such as affective, tactile, and spatial experiences and actions experienced by the body. There are seven factors (spatial [10 items], tactile [10 items], physical [9 items], kinesthetic [9 items], emotional [8 items], characteristic feature [9 items], and affective experiences [8 items]) on the MISAC (Narin, 2019). The items on the MISAC are rated on a 7-point Likert-type scale ($1 = \textit{Very vivid and clear as in reality}$; $7 = \textit{no image appeared in my mind}$). A high score on the MISAC indicates a high power of mental imagery. Notably, respondents require a maximum of 25 and an average of 15 minutes to answer the MISAC.

For example, an item from the spatial dimension can be given as “Imagine a café you often go to or your favorite café in your mind. How clear and vivid you can imagine these: (a) the location of tables, chairs, cash register...etc.”. Another item example regarding the physical dimension is as follows: “There are several actions/movements you experience using your body (e.g., arms, legs, and body). How clearly and vividly can you imagine when you think of yourself doing these movements? (a) Carrying a heavy load on your back”.

TCIA (*Test of Creative Imagery Abilities*) is a test developed by Jankowska and Karwowski (2015) to measure creative imagery abilities. It was utilized by the authors of the current study after adapting it to Turkish (see Narin, 2019). The test consists of seven incomplete figures. Participants are asked to verbally produce and describe several images evoking these figures. Then, they are expected to select the most original image from those they produce, draw it, and title it. Next, the drawings are evaluated in three dimensions: vividness, originality, and convertibility. Also, the highest score that can be obtained on the TCIA test is 21. The test was utilized to establish criteria-related validity evidence in this study.

The *Metaphoric Thinking Test* (MTT) consists of 10 concepts and three initial sentences. The test aims to measure the participants' ability to make sense of an image, create conceptual images, and establish a similarity relationship (see Narin, 2019). The participants are expected to select only three of the ten concepts provided to them in the test, create sentences using the selected concepts in a new and different way, and complete the incomplete initial sentences in a way that creates new meaning and context. The associated concepts and sentences based on these concepts are then evaluated in the context of creative thinking with a rubric prepared by the researcher according to three levels: non-creative (0 points), partially creative (1 point), and high-level creative (2 points). The highest score that can be obtained on the MTT is 12. The test was utilized to establish criteria-related validity evidence in this study.

2.3. Procedures

First, the normality assumptions for each item were checked by using the criteria of ± 2 for skewness and ± 7 for kurtosis coefficients (West et al., 1995). All the items had low percentages (<5%) for the missing values, and all met the normality assumptions. To check whether the seven-factorial structure of the MISAC fits our data, we carried out CFA using the *lavaan* package (Rosseel, 2012) in R (R Core Team, 2022). All analyses were conducted using a diagonally weighted least squares estimator. As an indicator of a good fit, values below .05 for the root mean square error of approximation (RMSEA) and values above .95 for the Tucker-Lewis Index (TLI) and the Comparative Fit Index (CFI) were considered (Yu, 2002).

2.3.1. Item Selection via ACO

After the model fit was guaranteed, we ran the ACO algorithm to shorten the MISAC for our data using the *ShortForm* package (Raborn & Leite, 2018). The ACO algorithm mimics ants' behaviors to establish the shortest route to a food source as a model for searching the model fit processes of structural equation modeling (Marcoulides & Drezner, 2003). The goal of this approach was to reach an optimal or near-optimal model with a fewer number of items. For this, an iterative process is started with the ACO by using several parameters (i.e., ants, evaporation, and steps) and criteria (i.e., model fit indices) until a specified convergence criterion is met (i.e., the number of iterations) (see Leite et al., 2008).

In this current study, we also chose the same values of the model fit statistics mentioned earlier to evaluate the quality of the shortened scales generated by the ACO. As ACO follows a heuristic approach for calculating the probabilities of items to be selected for the short form, ACO may generate different short forms in each run (Leite et al., 2008). Thus, in the item selection process, we attempted to shorten the MISAC by selecting four or five items for each factor with minor modifications to the tuning parameters (i.e., the number of ants, evaporation rate, and steps) as follows (Raborn & Leite, 2018, p. 10):

- i. ants = 120,
- ii. evaporation = .95 (i.e., the percentage of the pheromone retained after evaporation between iterations), and
- iii. steps = 20 (i.e., a numeric value that sets the rule for stopping, which is the number of ants in a row for which the model does not change).

The algorithm's computational process took approximately one hour to run with these parameters. The ACO algorithm was rerun 24 times to select optimal item candidates encompassing each factor's context and aligned with relevant theoretical representations. After each run, we identified frequently selected items for each factor.

Furthermore, after the 15th run, content experts identified 13 items to be excluded from the short form due to their content. Consequently, we omitted these 13 items from the algorithm for the remaining runs. Then, we thoroughly reviewed the top five item sets selected by the algorithm. Subsequently, we engaged in discussions regarding item coverage with two content experts. Finally, collaborating with these experts and authors, we collectively chose the most suitable version of the short form. The codes used in this study are available in [Appendix A](#).

2.3.2. Gathering validity and reliability evidence

The means and standard deviations were also calculated for each factor and item. In addition, we calculated both Cronbach's α and McDonald's (1999) ω as reliability evidence by using the *psych* package (Revelle, 2019). Notably, we considered ω and $\alpha > .70$ as a threshold for moderate reliability (Brunner et al., 2012; Nunnally & Bernstein, 1994), acknowledging the contextual considerations and potential trade-offs associated with reliability standards in research. The inter-correlations of the factors from the shortened MISAC with external criteria (i.e., metaphorical thinking and creativity imagery abilities) were calculated to gather concurrent validity evidence. The purpose of this analysis was to check whether the correlations obtained between the factors of the full scale and external variables were maintained within the shortened scale.

The ACO algorithm allows the selection of invariant items among specified groups, as demonstrated in various studies (Jankowsky et al., 2020; Olaru et al., 2019; Schroeders et al., 2016). However, due to our sample's gender imbalance (29.2% males) and relatively small male group size ($n = 145$), our initial analysis using modified functions from Jankowsky et al. (2020) and Olaru and Jankowsky (2022) showed consistent differences between groups in almost every selection. This finding indicated the necessity for freely estimating coefficients in each selection, undermining the ACO algorithm's optimization. Consequently, we could not employ

ACO for item selection based on measurement invariance. Therefore, we checked for measurement invariance across genders by utilizing the *lavaan* package as described in Bulut (2020) after the item selection process.

Regarding this validity evidence, we aimed to show that the shortened version of the MISAC was equally robust across gender groups. A stepwise procedure that started from the least restrictive model to the more restricted model (i.e., configural, metric, scalar, and strict invariance model, respectively) was adopted (see Van de Schoot et al., 2012). Furthermore, to test the measurement invariance, differences between the model fits previously evaluated with the same criteria and values of $\Delta\chi^2$ and ΔCFI were calculated. Chen’s rule was followed (i.e., the ΔCFI is $<.01$) (Chen, 2007) to control whether both models fit equally well statistically.

3. RESULTS

In this study, the results of the analysis conducted in the prior research by Narin (2019) were shared to prove that the shortened scale has similar psychometric features to the full scale. Therefore, the information in Tables 1, 2, and 3 regarding the full scales was obtained from Narin’s study (2019). Following the item selection process, the most optimal results were achieved by selecting four items from each factor, consistently chosen by ACO algorithms. The selected items from the shortened scale are provided in Appendix B. As shown in Table 1, the results of the CFA model of the shortened scale confirmed the hypothesized 7-factor model of the full scale and demonstrated a good fit. Furthermore, the model fit statistics were determined to be very similar to the full scale of the MISAC.

Table 1. Model fit statistics of the full and shortened scales of the MISAC.

Scales	χ^2	<i>df</i>	<i>p</i>	CFI	TLI	RMSEA	Lower	Upper
Full	9450.99	1953	$<.001$	0.98	0.98	0.01	0.00	0.02
Shortened	371.92	329	$<.001$	0.97	0.97	0.02	0.00	0.02

Means, standard deviations (SD), reliability coefficients, and zero-order correlations of the full and shortened scales of the MISAC are provided in Table 2. The reliabilities of the shortened scale were lower than the coefficients of the full scale; however, they were higher than .70 and ranged from .72 to .80 for all factors. Thus, the values were within an acceptable range.

Table 2. Means, SDs, reliability coefficients, and zero-order correlations of the full and shortened scales of the MISAC.

		<i>M</i>	<i>SD</i>	α	ω	1	2	3	4	5	6	7
Full scale	1. Spatial	5.60	1.09	.90	.92	1						
	2. Tactile	6.12	0.79	.87	.89	.42*	1					
	3. Physical	5.92	0.91	.88	.91	.49*	.53*	1				
	4. Kinesthetic	5.01	1.10	.86	.89	.40*	.36*	.47*	1			
	5. Emotional	5.43	1.18	.85	.89	.35*	.30*	.49*	.26*	1		
	6. Characteristic	5.60	0.98	.84	.88	.38*	.54*	.41*	.42*	.35*	1	
	7. Affective	5.78	0.96	.83	.88	.38*	.44*	.57*	.43*	.39*	.42*	1
Shortened	1. Spatial	5.79	1.14	.80	.80	1						
	2. Tactile	6.07	0.93	.72	.73	.44*	1					
	3. Physical	5.98	0.94	.74	.75	.46*	.50*	1				
	4. Kinesthetic	5.23	1.20	.78	.79	.36*	.39*	.38*	1			
	5. Emotional	5.40	1.28	.75	.76	.30*	.33*	.39*	.23*	1		
	6. Characteristic	5.58	1.12	.72	.72	.39*	.50*	.37*	.36*	.32*	1	
	7. Affective	5.74	1.14	.77	.78	.38*	.44*	.46*	.40*	.36*	.36*	1

Note: Inter-dimensional scale correlations within each form. * $p < .001$

As shown in Table 2, zero-order correlations between the factors of the shortened scale were similar to those between the factors of the full scale. To gather concurrent validity evidence, the correlation coefficients were calculated between the external variables (i.e., metaphorical thinking and creativity imagery abilities) and the factors of the shortened scale and compared with the result obtained from the full scale. As presented in Table 3, the direction and magnitude of these relationships in the full scale (computed using sum scores) were generally maintained within the shortened scale.

Table 3. Correlations between external variables and factors of the full and shortened scales of the MISAC.

Factors	Full scale ($N = 420$)		Shortened scale ($N = 500$)	
	MTT	TCIA	MTT	TCIA
Spatial	0.11*	0.01	0.10*	0.05
Tactile	0.18***	0.03	0.14**	0.09*
Physical	0.02	0.05	0.01	0.12**
Kinesthetic	0.11*	0.15**	0.06	0.12**
Emotional	0.01	0.01	0.09*	0.06
Characteristic	0.13*	0.08	0.12**	0.17***
Affective	0.11*	0.09	0.12**	0.16***

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

We tested the measurement invariance in the shortened scale to gather additional validity evidence and to examine whether this form of the scale maintained the same factorial structure across gender groups. Thus, the results of measurement invariance tests are presented in Table 4.

Table 4. Measurement invariance tests across gender.

Invariance test	χ^2	df	CFI	RMSEA	$\Delta\chi^2$	ΔCFI
Configural	350.68	658	.977	.015	-	-
Metric	433.49	679	.973	.015	24.212	.003
Scalar	455.67	700	.968	.017	34.301	.010
Partial scalar	433.05	699	.976	.014	5.4924	.003
Strict	461.25	727	.972	.015	38.095	.004

The first line of Table 4 shows the results of the baseline model (i.e., the model parameters are freely estimated across gender groups). These results were compared with later comparisons of more restrictive models. In the metric model, factor loadings were restricted to be equal for both genders. When the model fit values were compared, the chi-square difference test was not statistically significant ($\Delta\chi^2 = 24.212$, $df = 21$, $p = .28$), and ΔCFI was lower than 0.1, which indicated that the metric model fit the data equally across gender. As for the scalar model, the chi-square difference test was significant ($\Delta\chi^2 = 455.67$, $df = 21$, $p < .005$), and the change in CFI was also above an acceptable fit. Therefore, there was a lack of scalar invariance for the shortened scale. Thus, partial scalar invariance tests were established by freely estimating regression coefficients between the Physical Factor and item 24 (M24) for females and males. Then, it was indicated by the comparison of the adjusted scalar model and the metric model that partial scalar invariance was established for the scale ($\Delta\chi^2 = 5.4924$, $df = 20$, $p = .9$, $\Delta CFI < .1$). Finally, strict invariance was checked by using the adjusted scalar model. However, the residuals were constrained to be equal for females and males. Thus, it was shown in the results that the chi-square difference test was not significant ($\Delta\chi^2 = 38.095$, $df = 28$, $p = .9$) and the ΔCFI was lower than 0.1. As a result, strict invariance across females and males was

established. In sum, when these results were combined with moderate reliability and a good model fit, it was concluded that the shortened form had similar features to the full scale.

4. DISCUSSION and CONCLUSION

The primary objective of this current study was to develop a reliable and valid short form of the MISAC by utilizing the ACO algorithm. In addition, another aim was to gather validity and reliability evidence for the shortened version of the MISAC and test measurement invariance across gender groups. The ACO produced a 4-item per factor with a total of 28 items in the shortened scale (around 44% shorter). This finding suggests that responding to the shortened version of the MISAC can take approximately eight minutes on average. Hence, the shortened scale can allow researchers to collect data more flexibly and efficiently while reducing time, cost, and respondent burden (Basarkod et al., 2018).

Notably, the factorial structure and inter-correlations were maintained for the factors within the shortened scale. Furthermore, the shortened scale maintained the content representation across the seven factors underlying the MISAC. The item selection process inevitably involves a trade-off between their predictive strength and ensuring comprehensive content coverage (Leite et al., 2008; Raborn et al., 2019). Additionally, item sampling methods are closely connected to the specific elements within the construct being studied and the available item pool (Jankowsky et al., 2020). Given the relatively constrained size of the MISAC's original item pool, the ACO methodology adeptly extracted items that aligned statistically and conceptually with the intended content.

Our findings regarding the association between mental imagery and external variables (i.e., metaphorical thinking and creativity imagery abilities) were consistent with the results of the full scale. Obtaining the same results with the shortened version of the MISAC indicated that the short version has similar relationships with external variables, as seen in the full version. Research demonstrates that mental imagery serves as a foundational cognitive skill not only in creating mental representations of “images” but also in comprehending metaphors, thereby indicating its pivotal role in cognitive processes and creativity (Cornelissen & Clarke, 2010; Pérez-Fabello et al., 2016). Therefore, this finding holds crucial significance, indicating that scores derived from the short form effectively pinpoint the nuanced interplay between mental imagery and mentioned external variables. This validation solidifies the utility and applicability of the shortened MISAC in assessing and understanding the intricate cognitive mechanisms at play.

Smith et al. (2020) noted that shortening a scale brings several drawbacks. One such drawback related to reliability was evident in this study. With fewer items included in each factor, the shortened MISAC demonstrated only acceptable reliability. The measurement invariance results also revealed that strict invariance across females and males was attained using the adjusted scalar model. This conclusion stemmed from the observed disparity in the regression coefficient between item 24 (Sensing the texture of warm water) for females and males within the Physical Factor, suggesting varying interpretations of this item between genders. This discrepancy might be linked to gender's substantial influence on thermal perception (Schellen et al., 2013). The mental perception of warm water's temperature and texture may vary depending on gender. Hence, we recommend considering the shortened MISAC depending on the sample characteristics and research objectives. The original MISAC form might remain preferable when investigating gender differences.

The results showed that the ACO algorithm produced a shortened scale that satisfactorily showed good psychometric properties. Hence, the shortened scale can be considered a suitable alternative to the full scale in measuring mental imagery in the context of artistic creativity. The results of this current study are similar to previous studies that indicate that the ACO algorithm provides an effective procedure for shortening scales (Leite et al., 2008; Marcoulides & Drezner, 2003). Nevertheless, the ACO algorithm should be run multiple times to determine

the appropriate items for content representability, as the item selection process should not be based solely on the algorithms (Kleka & Soroko, 2018). Therefore, these automated algorithms may guide researchers in efficiently selecting their items (Yarkoni, 2010). The manual selection of items does have disadvantages and does not always offer an optimal solution (Olaru et al., 2015; Sandy et al., 2014). Thus, running automated algorithms and examining results in terms of relevant theories may be preferable.

Overall, researchers aiming to collect data regarding mental imagery may utilize the shortened version of the MISAC to save time while maintaining a high level of reliability, validity, and similar features to the full scale. So, researchers aiming to use scales that include items with relatively demanding cognitive loads, such as the MISAC, can follow similar procedures to obtain psychometrically sound brief scales.

Some methodological limitations in this study should be considered. First, there were limitations regarding the sample's representativeness, as it consisted solely of university students enrolled in undergraduate programs at art education institutions. Because of the gender imbalance in our sample and the limited number of male students, we were unable to utilize Jankowsky et al.'s (2020) and Olaru and Jankowsky's (2022) functions, which could have enabled us to select measurement invariance as a means to create a short form within the ACO algorithm. Therefore, for future research without these limitations, it is recommended that the functions be adapted to their specific dataset and the ACO algorithm employed accordingly.

Since this study was conducted within an art education group selected through a rigorous process, certain items might have been relatively effortless for participants to imagine mentally. To thoroughly investigate mental imagery within artistic work, other programs that require creative skills, such as design, architecture, and communication, should also be included in future research. Additionally, in this study, we could not consider students' academic year levels as a grouping variable due to the limited sample size in specific year cohorts. Future research could explore potential mean-level differences in students' abilities in mental imagery throughout their university education in the analysis.

In this study, there were no external variables that could reveal moderate or high correlations within our data set. Thus, future research can include additional variables to collect convergent or divergent validity. Exploring drawing skills, visual thinking abilities, and imaginative thinking skills through well-known assessments (e.g., The Torrance Tests of Creative Thinking) can be useful for gathering such evidence. Furthermore, latent group differences in mental imagery, in conjunction with these variables, can be examined while considering the previously mentioned grouping variables. Finally, since we aimed to shorten the scale, the reliability level decreased compared to the full scale. Therefore, using the shortened or full scale depends on the aim of future research. For example, suppose the plan is to utilize students' scores for individual-level decisions, such as selecting individuals for programs that require artistic skills or within the diagnostic processes for especially talented individuals. In that case, we recommend utilizing the full scale, as is strongly emphasized in other studies (e.g., Kruyen et al., 2014). However, the shortened scale is recommended if the aim is to analyze scores at a group level, such as modeling mental imagery or determining associations with other relevant variables. This approach saves time and reduces response bias during assessment sessions.

Acknowledgments

This paper was extracted from the first author's doctoral dissertation.

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(s). **Ethics Committee Number:** Cukurova University, Social Sciences and Humanities Research Ethics Committee, 03/04/2018-E.49944.

Contribution of Authors

Handan Narin Kızıltan: Investigation, Conception, Literature Review, Materials, Data Collection and Processing, Visualization, Writing. **Hatice Çiğdem Bulut:** Investigation, Conception, Literature Review, Design, Data Analysis, Visualization, Supervision, Writing.

Orcid

Handan Narin Kızıltan  <https://orcid.org/0000-0002-2164-8389>

Hatice Çiğdem Bulut  <https://orcid.org/0000-0003-2585-3686>

REFERENCES

- Andrade, J., May, J., Deeprose, C., Baugh, S.J., and Ganis, G. (2014). Assessing vividness of mental imagery: The Plymouth Sensory Imagery Questionnaire. *British Journal of Psychology*, 105, 547-563. <https://doi.org/10.1111/bjop.12050>
- Arshamian, A., & Larsson, M. (2014). Same but different: The case of olfactory imagery. *Frontiers Psychology*, 5, 34. <https://doi.org/10.3389/fpsyg.2014.00034>
- Basarkod, G., Sahdra, B., & Ciarrochi, J. (2018). Body image-acceptance and action questionnaire-5: An abbreviation using genetic algorithms. *Behavior Therapy*, 49(3), 388-402. <https://doi.org/10.1016/j.beth.2017.09.006>.
- Betts, G.H. (1909). *The distribution and functions of mental imagery*. New York, Columbia University.
- Bhattacharya, J., & Petsche, H. (2002). Shadows of artistry: cortical synchrony during perception and imagery of visual art. *Cognitive Brain Research*, 13(2), 179-186. [https://doi.org/10.1016/S0926-6410\(01\)00110-0](https://doi.org/10.1016/S0926-6410(01)00110-0)
- Botes, E., Dewaele, J.M., & Greiff, S. (2021). The development and validation of the short form of the foreign language enjoyment scale. *The Modern Language Journal*, 105(4), 858-876. <https://doi.org/10.1111/modl.12741>
- Bowns, R., Loeffelman, J.E., Steinley, D., & Sher, K.J. (2022). A brief young adult alcohol problems screening test: Short form development using combinatorics. *Journal of American College Health*, 1-7. <https://doi.org/10.1080/07448481.2022.2095870>
- Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, 80, 796-846. <https://doi.org/10.1111/j.1467-6494.2011.00749.x>
- Bulut, O. (2020). *Testing for Measurement Invariance in R*. Retrieved from <https://okan.cloud/posts/2020-12-21-testing-for-measurement-invariance-in-r/>
- Calabrese, L., & Marucci, F.S. (2006). The influence of expertise level on the visuo-spatial ability: Differences between experts and novices in imagery and drawing abilities. *Cognitive Processing*, 7(1), 118-120. <https://doi.org/10.1007/s10339-006-0094-2>.
- Carr, T., Moss, T., & Harry, D. (2005). The DAS24: A short form of the Derriford Appearance Scale DAS59 to measure individual responses to living with problems of appearance. *British Journal of Health Psychology*, 10(2), 285-298. <https://doi.org/10.1348/135910705X27613>
- Chamberlain, R., Drake, J.E., Kozbelt, A., Hickman, R., Siev, J., & Wagemans, J. (2019). Artists as experts in visual cognition: An update. *Psychology of Aesthetics, Creativity, and the Arts*, 13(1), 58. <https://doi.org/10.1037/aca0000156>
- Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14, 464-504. <https://doi.org/10.1080/10705510701301834>
- Cornelissen, J.P., & Clarke, J.S. (2010). Imagining and rationalizing opportunities: Inductive reasoning and the creation and justification of new ventures. *Academy of Management Review*, 35(4), 539-557. <https://doi.org/10.5465/amr.35.4.zok539>

- Cumming, J., & Eaves, D.L. (2018). The nature, measurement, and development of imagery ability. *Imagination, Cognition and Personality*, 37(4), 375-393. <https://doi.org/10.1177/0276236617752439>
- Deaver, S.P., & Shiflett, C. (2011). Art-based supervision techniques. *The Clinical Supervisor*, 30(2), 257-276. <https://doi.org/10.1080/07325223.2011.619456>
- Deneubourg, J.L., Pasteels, J.M., & Verhaeghe, J.C. (1983). Probabilistic behaviour in ants: A strategy of errors? *Journal of Theoretical Biology*, 105, 259-271. [https://doi.org/10.1016/S0022-5193\(83\)80007-1](https://doi.org/10.1016/S0022-5193(83)80007-1)
- Dilmac, O., & Kucuoglu, A. (2010). Güzel sanatlar eğitimi bölümleri resim-iş eğitimi anabilimdallarında sunum dosyasına dayalı özel yetenek sınav modeli [A sample of special ability test based on portfolio at the fine arts teaching departments]. *Journal of Institute of Fine Arts*, 18, 63-77.
- Dodson, B. (2013). *Keys to drawing with imagination: Strategies and exercises for gaining confidence and enhancing your creativity*. Pegasus Press.
- Dogan, B.G., Bulut, H.C. (2024) Abbreviation of parenting behaviors and temperament in children scales using genetic algorithms. *Current Psychology*, 43, 7044-7058. <https://doi.org/10.1007/s12144-023-04863-z>
- Drake, J.E., Simmons, S., Rouser, S., Poloes, I., & Winner, E. (2021). Artists excel on image activation but not image manipulation tasks. *Empirical Studies of the Arts*, 39(1), 3–16. <https://doi.org/10.1177/0276237419868941>
- Duncun, P. (2001). Visual culture: Developments, definitions, and directions for art education. *Studies in Art Education*, 42(2), 101-112. <https://doi.org/10.1080/00393541.2001.11651691>
- Efland, A.D. (2002). *Art and cognition: Integrating the visual arts in the curriculum*. Teachers College Press.
- Hetland, L., Winner, E., Veenema, S., & Sheridan, K.M. (2007). *Studio thinking: The real benefits of visual arts education*. Teachers College Press.
- Heid, K., Estabrook, M., & Nostrant, C. (2009). Dancing with line: Inquiry, democracy, and aesthetic development as an approach to art education. *International Journal of Education & the Arts*, 10(3).
- Holmes, E.A. & Mathew, A. (2010). Mental imagery in emotion and emotional disorders. *Clinical Psychology Review*, 30(3), 349-362. <https://doi.org/10.1016/j.cpr.2010.01.001>
- Jankowska, D.M., & Karwowski, M. (2015). Measuring creative imagery abilities. *Frontiers in Psychology*, 6, 1591. <https://doi.org/10.3389/fpsyg.2015.01591>
- Jankowska, D.M. & Karwowski, M. (2020). Mental imagery and creativity. <https://doi.org/10.31234/osf.io/eyfxr>
- Jankowsky, K., Olaru, G., & Schroeders, U. (2020). Compiling measurement invariant short scales in cross-cultural personality assessment using ant colony optimization. *European Journal of Personality*, 34(3), 470-485. <https://doi.org/10.1002/per.2260>
- Ji, J.L., Kavanagh, D.J., Holmes, E.A., MacLeod, C., & Di Simplicio, M. (2019). Mental imagery in psychiatry: Conceptual & clinical implications. *CNS spectrums*, 24(1), 114-126. <https://doi.org/10.1017/S1092852918001487>
- Kleka, P., & Soroko, E. (2018). How to avoid the sins of questionnaires abridgement? *Survey Research Methods*, 12(2), 147-160. <https://doi.org/10.18148/srm/2018.v12i2.7224>
- Kosslyn, S.M., Ganis, G., & Thompson, W.L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, 2, 635-642. <https://doi.org/10.1038/35090055>
- Kozhevnikov, M., Kozhevnikov, M., Yu, C.J., & Blazhenkova, O. (2013). Creativity, visualization abilities, and visual cognitive style. *British Journal of Educational Psychology*, 83, 196-209. <https://doi.org/10.1111/bjep.12013>.
- Krosnick, J.A., Holbrook, A.L., Berent, M.K., Carson, R.T., Hanemann, W.M., Kopp, R.J., ..., Conaway, M. (2002). The impact of “no opinion” response options on data quality: Non-

- attitude reduction or an invitation to satisfice? *The Public Opinion Quarterly*, 66(3), 371-403. <https://doi.org/10.1.1.141.7834>
- Kruyen, P.M., Emons, W.H., & Sijtsma, K. (2014). Assessing individual change using short tests and questionnaires. *Applied Psychological Measurement*, 38, 201-216. <https://doi.org/10.1177/0146621613510061>
- Lakoff, G., & Johnson, M. (2010). *Metaphors we live by* (G.Y. Demir, Trans.). (Rev. ed. 2nd ed.). Paradigma Press.
- Leite, W.L., Huang, I.C., & Marcoulides, G.A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, 43(3), 411-431. <https://doi.org/10.1080/00273170802285743>
- Marcoulides, G.A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(1), 154-164. https://doi.org/10.1207/S15328007SEM1001_8
- Marks, D.F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology*, 64, 17-24. <https://doi.org/10.1111/j.2044-8295.1973.tb01322.x>
- Marks, D.F. (1995). New directions for mental imagery research. *Journal of Mental Imagery*, 19(3-4), 153-167.
- McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- Miller, L.A. (2014). A self-report measure of cognitive processes associated with creativity. *Creativity Research Journal*, 26(2), 203-218. <https://doi.org/10.1080/10400419.2014.901088>.
- Narin, H. (2019). Sanat eğitimi öğrencilerinin zihinsel imgeleme kapasitesini belirlemeye yönelik bir ölçme aracı geliştirme çalışması [A study on developing a measurement tool for determining mental imagery capacity of art education students]. [Unpublished doctoral thesis]. Cukurova University.
- Niessen, A.S.M., Meijer, R.R., & Tendeiro, J.N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use?. *Journal of Research in Personality*, 63, 1-11. <https://doi.org/10.1016/j.jrp.2016.04.010>
- Nunnally, J.C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed). McGraw-Hill.
- O'Donoghue, D. (2011). Has the art college entry portfolio outlived its usefulness as a method of selecting students in an age of relational, collective and collaborative art practice? *International Journal of Education & the Arts*, 12(3), 1-27.
- Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, 59, 56-68. <https://doi.org/10.1016/j.jrp.2015.09.001>
- Olaru, G., & Jankowsky, K. (2022). The HEX-ACO-18: Developing an age-invariant HEXACO short scale using ant colony optimization. *Journal of Personality Assessment*, 104(4), 435-446. <https://doi.org/10.1080/00223891.2021.1934480>
- Ozmutlu, A., & Tomak, A. (2021). Examination of plastic arts field special ability exams in Turkey on the basis of their characteristics and risks. *International Journal of Interdisciplinary and Intercultural Art*, 6(12), 39-56.
- Palmiero, M., Piccardi, L., Nori, R., Palermo, L., Salvi, C., & Guariglia, C. (2016) Editorial: Creativity and Mental Imagery. *Frontiers Psychology*, 7(1280). <https://doi.org/10.3389/fpsyg.2016.01280>
- Park, M., & Yoo, J. (2020). Effects of perceived interactivity of augmented reality on consumer responses: A mental imagery perspective. *Journal of Retailing and Consumer Services*, 52, 101912. <https://doi.org/10.1016/j.jretconser.2019.101912>
- Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10), 624-634. <https://doi.org/10.1038/s41583-019-0202-9>

- Pearson, J., Naselaris, T., Holmes, E.A., & Kosslyn, S.M. (2015). Mental imagery: Functional mechanisms and clinical applications. *Trends in Cognitive Sciences*, 19(10), 590-602. <https://doi.org/10.1016/j.tics.2015.08.003>.
- Pérez-Fabello, M.J., & Campos, A. (2007). Influence of training in artistic skills on mental imaging capacity, *Creativity Research Journal*, 19(2-3), 227-232.
- Pérez-Fabello, M.J., Campos, A., & Campos-Juanatey, D. (2016). Is object imagery central to artistic performance?. *Thinking Skills and Creativity*, 21, 67-74. <https://doi.org/10.1016/j.tsc.2016.05.006>.
- R Core Team (2022). *R: A Language and environment for statistical computing*. (Version 4.0) [Computer software]. <https://cran.r-project.org>
- Raborn, A.W., & Leite, W.L. (2018). ShortForm: An R package to select scale short forms with the ant colony optimization algorithm. *Applied Psychological Measurement*, 42(6), 516-517. <https://doi.org/10.1177/0146621617752993>
- Raborn, A.W., Leite, W.L., & Marcoulides, K.M. (2020). A comparison of metaheuristic optimization algorithms for scale short-form development. *Educational and Psychological Measurement*, 80(5), 1-22. <https://doi.org/10.1177/0013164420906600>
- Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences*, 35(4), 212-220. <https://doi.org/10.1027/1614-0001/a000141>
- Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. [R package]. Retrieved from <https://cran.r-project.org/package=psych>.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Rosenberg, H.S. (1987). Visual artists and imagery. *Imagination, Cognition and Personality*, 7(1), 77-93. <https://doi.org/10.2190/AVJ5-N24B-P7MC-HR4R>.
- Sandy, C.J., Gosling, S.D., & Koelkebeck, T. (2014). Psychometric comparison of automated versus rational methods of scale abbreviation: An illustration using a brief measure of values. *Journal of Individual Differences*, 35(4), 221-235. <https://doi.org/10.1027/1614-0001/a000144>
- Saulsman, L.M., Ji, J.L., & McEvoy, P.M. (2019). The essential role of mental imagery in cognitive behaviour therapy: What is old is new again. *Australian Psychologist*, 54(4), 237-244. <https://doi.org/10.1111/ap.12406>
- Schellen, L., Loomans, M., de Wit, M., & van Marken Lichtenbelt, W. (2013). The influence of different cooling techniques and gender on thermal perception. *Building Research & Information*, 41(3), 330-341. <https://doi.org/10.1080/09613218.2013.772002>
- Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PloS One*, 11(11), e0167110. <https://doi.org/10.1371/journal.pone.0167110>
- Schwarz, S., Grasmann, D., Schreiber, F., & Stangier, U. (2020). Mental imagery and its relevance for psychopathology and psychological treatment in children and adolescents: A systematic review. *International Journal of Cognitive Therapy*, 13, 303-327. <https://doi.org/10.1007/s41811-020-00092-5>
- Serig, D. (2006). A conceptual structure of visual metaphor. *Studies in Art Education*, 47(3), 229-247. <https://doi.org/10.1080/00393541.2006.11650084>
- Sheehan, P.W. (1967). A shortened form of Betts' Questionnaire Upon Mental Imagery. *Journal of Clinical Psychology*, 23(3), 386-389. [https://doi.org/10.1002/1097-4679\(196707\)23:3<386::AID-JCLP2270230328>3.0.CO;2-S](https://doi.org/10.1002/1097-4679(196707)23:3<386::AID-JCLP2270230328>3.0.CO;2-S)
- Smith, G.T., McCarthy, D.M., & Anderson, K.G. (2000). On the sins of short-form development. *Psychological Assessment*, 12(1), 102. <https://doi.org/10.1037/1040-3590.12.1.102>

- Swindle, R., Cameron, A., & Rosen, R. (2006). A 15-item short form of the psychological and interpersonal relationship scales. *International Journal of Impotence Research*, 18(1), 82–88. <https://doi.org/10.1038/sj.ijir.3901381>
- Taskesen, S. (2019). An investigation on special talent exams in the division of art teaching. *Journal of Education and Training Studies*, 7(10), 86-97. <https://doi.org/10.11114/jets.v7i10S.4554>
- Tay, J. (2019). *Art teachers' perceptions about visual arts giftedness: Content and construct validation of perceptions about art giftedness* [Doctoral dissertation, Purdue University]. ProQuest Dissertations Publishing.
- Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4), 486-92. <https://doi.org/10.1080/17405629.2012.686740>
- Vellera, C., & Gavard-Perret, M. L. (2012). *Is mental imagery ability an element for identifying creative consumers?*. <https://halshs.archives-ouvertes.fr/halshs-00851322>
- Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3), 363-373. <https://doi.org/10.1016/j.jrp.2010.04.001>
- Yilmaz, S. (2016). The relation between visual perception tests organized for special talent exams and academic achievement in art teaching program. *The Black Sea Journal of Social Sciences*, 8(15), 55-74.
- Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* [Doctoral dissertation, University of California]. ProQuest Dissertations Publishing.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3), 236-247. <https://doi.org/10.1016/j.ijresmar.2010.02.004>
- West, S.G., Finch, J.F., Curran, P.J., & Hoyle, R.H. (1995). Structural equation modeling: Concepts, issues, and applications. In Hoyle, R.H (Ed.). *Structural equation models with nonnormal variables: Problems and remedies*, (p. 55-75). Sage
- Ziss, A. (2011). *The science of artistic assimilation of aesthetic reality* (2nd ed.). (Y. Şahan, Trans.). Hayalbaz Books.

6. APPENDIX

6.1. Appendix A. Utilized Code for Running the ACO Algorithm

```

# Load packages
library(ShortForm)
library(lavaan)

# Load data
misac <- read_excel("C:/.../Research/SA/misac.xlsx")
misac_v1 <- data.matrix(misac[,6:68])

# Run the ACO logarithm
misac_short <- antcolony.lavaan(data = misac_v1,
                               ants = 120,
                               evaporation = 0.95,
                               antModel = 'char =~ M6+ M2+ M8+ M7+ M10+ M3+ M1
                                           kine =~ M16+ M20+ M18+ M14+ M15+ M12
                                           tact =~ M22+ M26+ M21+ M29+ M23+ M27+ M24+ M25+ M30+ M28
                                           spat =~ M43+ M45+ M48+ M41+ M44
                                           phys =~ M51+ M54+ M52+ M55+ M56+ M59+ M53+ M58
                                           emot =~ M69+ M67+ M66+ M65+ M63+ M70+ M64
                                           affe =~ M75+ M72+ M76+ M80+ M73+ M74+ M79',
                               list.items = list(c('M6', 'M2', 'M8', 'M7', 'M10', 'M3', 'M1'),
                                                  c('M16', 'M20', 'M18', 'M14', 'M15', 'M12'),
                                                  c('M22', 'M26', 'M21', 'M29', 'M23', 'M27', 'M24', 'M25', 'M30', 'M28'),
                                                  c('M43', 'M45', 'M48', 'M41', 'M44'),
                                                  c('M51', 'M54', 'M52', 'M55', 'M56', 'M59', 'M53', 'M58'),
                                                  c('M69', 'M67', 'M66', 'M65', 'M63', 'M70', 'M64'),
                                                  c('M75', 'M72', 'M76', 'M80', 'M73', 'M74', 'M79')),
                               full = 50, i.per.f = c(4,4,4,4,4,4,4),
                               factors = c('char', 'kine', 'tact', 'spat', 'phys', 'emot', 'affe'),
                               steps = 20,
                               fit.indices = c('cfi', 'rmsea'),
                               fit.statistics.test = "(cfi > 0.95)&(rmsea < 0.05)",
                               summaryfile = 'summary.txt',
                               feedbackfile = 'iteration.html',
                               max.run = 1000)

# print selected items
misac_short$best.syntax

```

6.1. Appendix B. Results of confirmatory factor analysis of the shortened scale

Factor	Item	Rephrased Item Labels and prompts	Estimate (SE)
		How vividly can you imagine your favorite café in your mind?	
Spatial	M43	Visualizing the interior dimensions of the cafe	.67 (.06) ^{***}
	M45	Visualizing the placement of tables, chairs, cash register, etc. in the cafe	.80 (.06) ^{***}
	M44	Visualizing the height of the cafe's ceiling	.58 (.07) ^{***}
	M41	Visualizing the color and shape of the cafe's signboard	.77 (.06) ^{***}
		How clearly can you imagine the sensations you feel with your hands?	
Tactile	M22	Sensing the texture of cotton	.64 (.05) ^{***}
	M23	Sensing the texture of a thorn	.62 (.06) ^{***}
	M24	Sensing the texture of warm water	.57 (.07) ^{***}
	M28	Sensing the texture of silk fabric	.69 (.05) ^{***}
		How vividly can you imagine these movements?	
Physical	M51	Walking uphill	.66 (.04) ^{***}
	M56	Carrying a heavy load on your back	.71 (.05) ^{***}
	M59	Throwing a basketball	.65 (.06) ^{***}
	M53	Climbing a tree	.59 (.07) ^{***}
		How clearly can you see various movements and situations related to a motorcycle and its actions?	
Kinesthetic	M16	Overcoming a bump/obstacle on a motorcycle	.58 (.06) ^{***}
	M20	Dragging a fallen motorcycle on the ground	.75 (.07) ^{***}
	M18	Motorcycle colliding rapidly with a vehicle	.79 (.06) ^{***}
	M14	Motorcycle swiftly passing by	.66 (.07) ^{***}
		How vividly can you imagine a feeling or emotion?	
Emotional	M69	Feeling guilt	.69 (.08) ^{***}
	M67	Experiencing panic/shock	.69 (.07) ^{***}
	M66	Feeling doubt	.74 (.08) ^{***}
	M70	Expressing astonishment	.53 (.09) ^{***}
		How vividly can you see a familiar friend in your mind?	
Characteristic	M7	Appearance while expressing joy	.60 (.07) ^{***}
	M8	Appearance when angered	.52 (.07) ^{***}
	M10	Notable behavior while eating (e.g., eating habits)	.72 (.07) ^{***}
	M3	Notable behavior while walking/stepping (e.g., stride length)	.67 (.07) ^{***}
		How clearly can you imagine the expressions or emotions?	
Affective	M75	A cat with a full stomach	.68 (.06) ^{***}
	M80	A dog growling upon seeing a stranger	.70 (.07) ^{***}
	M73	Eating situation of a child with a sore throat	.77 (.06) ^{***}
	M74	Body of a sleep-deprived person	.59 (.06) ^{***}

Note. Total explained variance ($R^2 = 59\%$), * $p < .05$, ** $p < .01$, *** $p < .001$

The use of ChatGPT in assessment

Mehmet Kanik ^{1*}

¹Final International University, Faculty of Educational Sciences, English Language Teaching Program, Girne, North Cyprus

ARTICLE HISTORY

Received: Oct. 22, 2023

Accepted: Aug. 12, 2024

Keywords:

ChatGPT,
AI,
Assessment,
Item-generation,
Item analysis.

Abstract: ChatGPT has surged interest to cause people to look for its use in different tasks. However, before allowing it to replace humans, its capabilities should be investigated. As ChatGPT has potential for use in testing and assessment, this study aims to investigate the questions generated by ChatGPT by comparing them to those written by a course instructor. To investigate this issue, this study involved 36 junior students who took a practice test including 20 multiple-choice items generated by ChatGPT and 20 others by the course instructor, resulting in a 40-item test. Results indicate that there was an acceptable degree of consistency between the ChatGPT and the course instructor. Post-hoc analyses point to consistency between the instructor and the chatbot in item difficulty, yet the chatbot's results were weaker in item discrimination power and distractor analysis. This indicates that ChatGPT can potentially generate multiple-choice exams similar to those of the course instructor.

1. INTRODUCTION

Swiecki et al. (2022) criticize standard assessment paradigms for being onerous, discrete, uniform, antiquated, and lacking authenticity. They propose that artificial intelligence (AI) can offer solutions to these challenges. In a review article on the use of AI in student assessment, González-Calatayud et al. (2021) argue that AI technologies remain underutilized in education due to users' lack of knowledge. However, within the past few years, there have been discussions on the impact of AI language models with the emergence of ChatGPT, a chatbot released by a company named OpenAI (chat.openai.com). This interest has also led to a surge in research studies in education, primarily focusing on language learning (Crompton & Burke, 2023).

Nevertheless, ChatGPT came with concerns and controversies, especially within the field of education. One of the initial reactions was of the negative kind as reports revealed that students had ChatGPT or other AI models to write projects and homework assignments for them. However, these language models may also offer some potential benefits and uses. For instance, Okonkwo and Ade-Ibijola (2021) identified several possible uses of chatbots in education including teaching, learning, and assessment. In Crompton and Burke's recent review (2023), themes such as assessment/evaluation, prediction, AI assistance, intelligent tutoring systems,

*CONTACT: Mehmet KANIK ✉ mehmetkanik@gmail.com 📍 Final International University, Faculty of Educational Sciences, English Language Teaching Program, Girne, North Cyprus

© The Author(s) 2024. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

and student learning management emerged as common applications of AI in education. This underscores the potential of AI language models like ChatGPT in education. Yet, before making use of such technologies, it is crucial to scrutinize their use, supported by evidence, as they may not always produce satisfactory or accurate content (van Dis et al., 2023). Therefore, this study attempts to investigate the use of ChatGPT in test preparation and assessment.

1.1. Literature Review

Gardner et al. (2021) state that Page (1966) “foresaw a time in the future when natural language processing (NLP) would achieve the technical maturity to enable machines to learn and understand how to assess the existence of the many complex trins in human writing” (p. 1208). Gardner et al. (2021) elaborate on this idea, asserting that machines can assess students on their knowledge of the content if the machine is trained on that content and trained to ask questions. To some extent, Page’s (1966) prediction has become a reality as AI technologies now possess such capabilities. They can even do more. ChatGPT, for instance, has great capabilities that can contribute to teaching and assessment in a variety of ways. In an article, for instance, Lo (2023) reviewed studies on ChatGPT and identified five key uses of it in teaching and assessment, ranging from generating course materials to performing language translation. Lo (2023) also suggests that students can use it in preparing writing assignments for assessment. They can draft papers and have ChatGPT evaluate them for errors, and then the students can finalize their papers. As such, the chatbot can act as a useful scaffolding tool. According to this review, instructors can have it generate assessment tasks and evaluate student performance. Assessment and evaluation emerged as the most common use of AI technologies in higher education, as revealed by Crompton and Burke’s (2023) review, encompassing automatic assessment, test generation, feedback, online activity review, and the evaluation of educational resources. Formative assessment, automated scoring, and comparisons between AI and non-AI assessment methods are also central to the research on assessment (González-Calatayud et al., 2021).

In a more detailed look at the contributions AI can make to overcome the problems in the standard assessment paradigm, Swiecki et al. (2022) suggest such uses as automated assessment construction, AI-assisted peer assessment, writing analytics, electronic assessment platforms, stealth assessment, latent knowledge estimation, learning processes, computerized adaptive testing, virtual simulations to add authenticity and modernized digital assessment by incorporating computational media such as AI-supported word processing. As an AI tool, Halaweh (2023) highlights the time and effort ChatGPT helps save and compares it to other tools like search engines and spreadsheets that are used to help with searching for information, calculations, and organizing data without concern, which were tasks that people had to do without the assistance of technology. The researcher suggests that as there are no concerns with using these tools so should there be no concern with using ChatGPT’s abilities to produce and edit texts by considering it as a tool to save time and effort.

Yet, there are obvious concerns about the ethicality of using ChatGPT as it is capable of producing texts quickly and can cause ethical issues when used to replace one’s role as the writer of a text. Dowling and Lucey (2023) found, for example, that ChatGPT can produce articles that can go through a peer-review process as the three articles produced by ChatGPT got high ratings from the reviewers. If the authorship is falsely claimed, they suggest, then ethical issues ensue. ChatGPT poses some issues for the users as well. For example, it can rely on biased data, not having up-to-date information, and generate incorrect or fake information. It can also present issues to educators related to ethical concerns. It can lead students to be involved in plagiarism and have them bypass plagiarism detectors (Lo, 2023, p. 8). Mhlanga (2023), thus, suggests responsible and ethical uses of ChatGPT in education by highlighting factors including responsible AI use and educating students about it and its limitations, transparency in the use of ChatGPT, respect for privacy, accuracy of information, and the like.

Lo (2023) suggests that instructors can benefit from using ChatGPT as a valuable resource, as it helps in crafting course syllabi, teaching materials, and assessment tasks as long as issues related to the accuracy of the generated content are addressed. Al-Worafi et al. (2023) tried the feasibility of using ChatGPT for designing curriculum and syllabus, course content preparation, and writing exams. The chatbot got expert ratings from 50% to 92%. Overall, it could be suggested that ChatGPT can be a useful tool. One aspect that the researchers looked at was exam preparation and found that ChatGPT can be used for that purpose. The expert rating of appropriateness and accuracy of what ChatGPT produced was 70%. They caution, however, that the exams did not include all the learning outcomes. Other AI tools were used in studies to generate cloze tests and found that AI tools can enhance learning (Olney et al., 2017; Yang et al., 2021).

Regarding exam generation, Chen et al. (2018) mention two methods, rule-based and data-driven, used in automatic question generation, creating strong potential for AI use in education. They suggest that the rule-based method is prone to be influenced by the quality and quantity of rules developed by humans, which will be dependent on their knowledge, experience, and effort. They suggest, as an alternative, the use of data-driven methods which will not be dependent on human-generated rules. Their research with a data-driven method indicates the data set can affect the extent to which automatic question-generation methods can write quality items as their research shows that automatic question-generation methods did not perform well in a comprehensive data set.

Another aspect AI language models were used for was the a priori evaluation of the quality of the exams generated by humans. For example, Moore et al. (2022) utilized GPT-3 to evaluate the quality of the student-generated short-answer questions. Although their focus was on the extent to which students are able to generate quality test items, the results also indicated the use of GPT-3 in evaluating and assessing the content of students' work. They found, however, that GPT-3 matched human evaluation only for 40% of the questions. For the AI model, most of the questions were high quality as opposed to human experts who classified 68% of the questions as low quality. For GPT-3 this figure was only 9%. The researchers conclude that GPT-3 overestimated the quality of the questions. In assigning the items to the levels of Bloom's taxonomy, there was a disagreement between GPT-3 and human experts in 68% of the questions.

In another study, however, Moore et al. (2023) utilized GPT-4 along with human and automated rule-based methods in evaluating the quality of items by identifying item-writing flaws in multiple-choice items. They found that GPT-4 was able to identify 79% of the flaws identified by human annotators and matched 62% of the human quality evaluations. This may indicate that the more advanced language models become, the better they can perform pedagogical tasks, approximating the performance of experts. AI technologies have also been used in automated essay scoring and have been utilized commercially and in computerized adaptive testing both used commercially by testing companies like Pearson or ETS (Gardner et al., 2021). Thus, AI-based tools can automate traditional assessment by creating tests and automatically scoring them, eliminating some of the burden (Swiecki et al., 2022). Swiecki et al. (2022) list some challenges of AI-based assessment tools. They caution against directly accepting machine decisions and giving the responsibility to engineers with no contact with the students also causing a removal of accountability. They are also skeptical about eliminating the pedagogical role of assessment teachers may use to affect the teaching-learning process, limiting the process-based performance assessment. Another issue they raise is the data collection by AI technologies. These are quite valid concerns about AI-based technologies.

If AI tools like ChatGPT are used for creating exams or writing assessment tasks, exam validity and reliability become another concern because they are required qualities of any test (Thorndike & Thorndike-Christ, 2014). "Validity has to do with the degree to which test scores

provide information that is relevant to the inferences that are to be made from them” (Thorndike & Thorndike-Christ, 2014, p. 76) or put simply, measuring what we want to measure with it and usually focuses on content-, criterion-, and construct-related validity. Content validity is usually achieved by having an exam blueprint, or a table of specifications, which shows the content areas and cognitive processes involved and their respective weight in the test. Criterion and construct validation techniques may need correlation with other tests (Miller et al., 2013; Reynolds et al., 2009; Thorndike & Thorndike-Christ, 2014).

Reliability is defined as “the accuracy or precision of a measurement procedure” (Thorndike & Thorndike-Christ, 2014, p. 75) or “consistency or stability of assessment results” (Reynolds et al., 2009, p. 91) and it is essential for testing because the purpose of assessment is to make educational decisions and if the information to base the decisions on is not reliable, then the decisions are unlikely to be valid decisions (Reynolds et al., 2009; Thorndike & Thorndike-Christ, 2014) and essentially the test tests “nothing” (Thompson & Vacha-Haase, 2018, p. 231). The reliability of exams is usually measured by calculating a reliability coefficient by correlating the results of the same tests administered at different times, parallel forms of a test, two halves of a test, and scores awarded by different examiners (Reynolds et al., 2009). In all these approaches, the consistency between two sets of scores is at the focal point of measurement.

González-Calatayud et al. (2021) highlight that AI is mostly used for formative assessment. There do not seem to be studies focusing on its use in summative assessment by testing the applicability of tests generated by AI language models. To approach the issue more systematically, this study aims to analyze the results of an exam prepared by ChatGPT in tandem with the course instructor to better answer the question of whether ChatGPT can be used in test preparation by course instructors by running comparative post-hoc evaluations like reliability, item difficulty and discriminating power.

2. METHOD

This study employs a case study approach, incorporating both quantitative and qualitative research methods to provide a more in-depth analysis of the subject. To explore the quantitative aspect, correlation, paired-samples t-test and post-hoc analysis were utilized to examine the reliability between two tests and to examine various aspects of the test results. Qualitative data were also gathered and analyzed using content analysis. By combining quantitative correlational analysis with qualitative content analysis, this study aims to offer a rich, nuanced understanding of the case.

2.1. Context

The study was conducted at a private university in North Cyprus, which is an international university with a majority of international student population. The university has a faculty of educational sciences with both Turkish-medium and English-medium programs. English Language Teaching program, as well as all the other programs of the faculty, has a course on measurement and evaluation in education aiming to train student teachers on assessment and testing practices. The study is conducted within this class.

2.2. Participants

The participants were students enrolled in the said measurement and evaluation class offered as part of an undergraduate program in English Language Teaching. The class is a mandatory faculty class that all registered students should take. There were 44 students enrolled in the class. Thirty-six of them participated in the study by taking the review exam. One paper was eliminated for being incomplete as the student answered questions in one part of the exam which was mainly the instructor’s questions and did not complete most of the questions written by ChatGPT. The participant profile is outlined in [Table 1](#) below. The students come from Ivory Coast, Libya, North Cyprus, Russia, Türkiye, Turkmenistan, and Uzbekistan. Eighteen of the

participants were female while 17 were male. The mean age was 22.94, ranging from 21 and 28.

Table 1. *Participants.*

Nationality	N (35)	Age		Gender	
Türkiye	14	Mean	22.94	Female	18
Ivory Coast	7	Range	21-28	Male	17
Uzbekistan	6				
Libya	2				
North Cyprus	2				
Russia	2				
Turkmenistan	2				

2.3. Procedures

For the purpose of the study, a table of specifications including the content and learning outcomes of the said class was prepared. The table of specifications included 20 items distributed over the content of the class covered between the midterm exam and the final exam of the class. The same table of specifications was used also for the final exam of the class. The instructor of the class wrote 20 questions matching this table of specifications to ensure content validity. Then, the instructor pasted the content of the class lecture presentations into ChatGPT and asked the chatbot to write questions. A sample entry used, for example, reads “Using the following information, prepare a multiple-choice item on item analysis at Bloom’s knowledge level”. After ChatGPT produced 20 questions matching the same specifications, two sets were put together resulting in a 40-question multiple-choice test. Half of the students began with the instructor’s questions, while the other half started answering the questions written by ChatGPT. Students were also asked to write their comments on the questions for their perception of the test and the questions.

After the administration of the test, each exam paper was given several scores: One total score, one score for the questions by the instructor, one score for the questions written by ChatGPT, two scores each for the odd and even-numbered questions written by the instructor, ChatGPT and combined total resulting in nine different scores. These scores were put into statistical software for analysis. The main methods of statistical analyses were correlation and reliability analysis. Item analysis procedures were also conducted for item difficulty, item discrimination power, and distractor effectiveness. Students’ comments were analyzed qualitatively.

3. FINDINGS

The first analysis was calculating the internal reliability of the exam as well as the sections written by the instructor and ChatGPT. To calculate the internal consistency, the odd-numbered questions and the even-numbered questions were scored separately. This was done for the instructors’ and ChatGPT’s questions as well. The results of the analysis for the whole test yielded a score of .743, which is an acceptable value (Reynolds et al., 2009) as indicated in [Table 2](#).

Table 2. *Split-half reliability analysis results.*

Test	N	Odd M (SD)	Even M (SD)	Spearman-Brown coefficient
Instructor’s test	35	14.36 (4.59)	13.57 (5.60)	.636
ChatGPT’s test	35	15.64 (3.94)	16.07 (4.21)	.636
Combined	35	30.00 (7.52)	29.57 (8.47)	.743

Split-half reliability analysis was also calculated for the instructor's test and ChatGPT's tests using the Spearman-Brown formula. The obtained coefficient for both the instructor's and ChatGPT's questions was .636, indicating a moderate internal consistency, understandably a bit lower than the combined test since the sample size goes down in split-half analysis and lower reliability coefficients can be acceptable for short tests (McCowan & McCowan, 1999). Table 2 shows these results.

After establishing an acceptable degree of internal consistency, parallel forms reliability was calculated between the instructor's test and ChatGPT's test. The coefficient calculated for the reliability between these two forms was .80, which points to a good degree of consistency. This finding is important as it indicates that ChatGPT can prepare tests that function parallel to a course instructor's test. The results are depicted in Table 3.

Table 3. Consistency between the instructor's test and that of ChatGPT.

Test	N	M (SD)	Spearman-Brown coefficient
Instructor's half	35	28.00 (8.71)	.80
ChatGPT's half	35	31.57 (6.91)	

The question of whether the instructor in question writes consistent exams is a question in point here. To establish that, the reliability between the instructor's two tests given at two different times of the semester was calculated. A reliability coefficient of .92 was achieved, indicating a good degree of reliability, as shown in Table 4

Table 4. Reliability between two tests written by the course instructor.

Test	N	M (SD)	Cronbach's Alpha
Test 1	35	68.17 (16.36)	.92
Test 2	35	55.71 (18.98)	

ChatGPT's ability to write tests consistent with the course instructor's tests indicates its utility in helping with testing and assessment. The mean scores of the tests indicate, however, that the instructor's version may have been more challenging. The paired sample t-test was run, and the results showed that the students achieved higher scores in ChatGPT's set (M=31.57, SD= 6.91) than in the instructor's set (M=28, SD=8.71), resulting in a significant difference as can be seen Table 5.

Table 5. Paired samples t-test statistics.

Test	N	M (SD)	t	df	p
Instructor's set	35	28.00 (8.71)	-3.204	34	.003
ChatGPT's set	35	31.57 (6.91)			

After the reliability analyses, item analysis procedures were followed to see if ChatGPT writes items with a good level of difficulty and discrimination power.

3.1. The Results of Item Analysis

The difficulty index of the items demonstrates similar results from the instructor's and ChatGPT's sets. As Table 6 depicts, 70% of the instructor's test items proved to have moderate levels of difficulty while 65% of ChatGPT's test items fell into the moderate difficulty range. Both sets of test items had two that were identified as difficult. In terms of the easy items, 20% of the instructor's and 25% of ChatGPT's items were in the easy range. These results indicate

that both the course instructor and ChatGPT write questions at a comparable degree of difficulty.

Table 6. *Difficulty index.*

	Instructor	ChatGPT
Easy	4 (20%)	5 (25%)
Moderate	14 (70%)	13 (65%)
Difficult	2 (10%)	2 (10%)

Another relevant analysis is the discrimination power of the items (Ebel & Frisbie, 1986). In this analysis, the ratio of the correct answers by lower achieving to those of the higher achieving students is calculated. The results indicate that 75% of the instructor’s items are very good or reasonably good while only 50% of the items written by ChatGPT were good in terms of discrimination power. This indicates that ChatGPT fails to write items that can distinguish between the higher and the lower-achieving students. [Table 7](#) outlines these results.

Table 7. *Discrimination index.*

	Instructor	ChatGPT
Very good	11 (55%)	8 (40%)
Reasonably good	4 (20%)	2 (10%)
Marginal item	3 (15%)	6 (30%)
Poor item	2 (10%)	4 (20%)

3.2. Distractor Analysis

For the 20 four-option test items, both the instructor and ChatGPT wrote 60 distractors in total. The expectation for the distractors is that they are to be selected by some students and selected by the low-achieving students more than the high-achieving students (Miller et al., 2013). According to the results of the analysis, 90% of the distractors written by the instructor were selected by some students, while only 80% of those written by ChatGPT were selected by some students. The number of the instructor’s distractors that were selected by the lower group of students is 41, accounting for 71.6% of the total distractors. While this value is 34 for ChatGPT accounting for 56.6% of the distractors it wrote. In other words, 43.4% of the distractors written by ChatGPT were poor distractors as opposed to 28.4% of the instructor as shown in [Table 8](#). This can indicate that ChatGPT may not be apt to write plausible distractors.

On the other hand, in this specific case, the instructor may sometimes be writing distractors that may be confusing, as 10% of the distractors were selected more by the upper group, which indicates an issue. On the other hand, only one distractor written by ChatGPT was selected by the upper group more than the lower group. Thus, ChatGPT may be clearer in writing distractors although it may not always write plausible distractors.

Table 8. *Distractor analysis.*

	Functions as intended	Selected by none	Selected by the upper group more	Selected equally by upper and lower group
Instructor	43 (71.6%)	6 (10%)	6 (10%)	5 (8.3%)
ChatGPT	34 (56.6%)	12 (20%)	1 (1.66%)	13 (21.6%)

3.3. The Results of Qualitative Data Analysis

Students were asked to share their perceptions of the question in two sets briefly. The answers were not rich in that sense. Although they “did not see a big difference between them,” students had conflicting perceptions of the instructor’s and ChatGPT’s test items in some respects. One such perception is about the difficulty of the item sets. It seems that students related to the questions differently as some found the instructor’s items more difficult while some others thought the opposite as evident from the following samples on the instructor’s and ChatGPT’s test items respectively.

This part was harder than the other one.

I think questions are same but difficulty of questions got higher in this section.

Another issue is with the clarity of the questions. Some students did not find the instructor’s set clear while it was the opposite for some others. For example, one student said on the instructor’s items:

There are some questions which are unclear. Seemingly there are two correct answers in one question.

Commenting on ChatGPT’s items, on the other hand, students said the following:

Some of the questions were longer and were a bit harder to understand.

Questions are more complicated and confusing but the rest are easier. Questions are too long and also options. That’s why it is confusing.

Conversely, for some other students, “the questions are great. They are easy to understand and clear.” As indicated by the quotations above, the students found that ChatGPT’s items were longer, which they believed made them more confusing and difficult. Yet, when the length of the stem and alternatives in the number of words are considered, the data does not support this perception as the average length of the stems in the instructor’s set is 16.65 words, while it is 13.9 words for ChatGPT. When it comes to the length of the alternatives, however, ChatGPT wrote slightly longer alternatives as the length is 6.56 words for ChatGPT, whereas it is 6.12 words for the instructor. However, this difference is not large.

One other factor that some students highlighted was one related to vocabulary. For some students, ChatGPT’s questions included more unfamiliar words, making the test more challenging as indicated by the following remark.

It was much more difficult than the other 20 questions. It wants more information. There are so many words I did not know.

Finally, students also had conflicting ideas about the aim of the questions written by the instructor and ChatGPT. For some of them, the instructor’s questions asked for general information while ChatGPT asked for specific information. It was the opposite for some other students. For example, these are comments about the instructor’s set.

The questions were mostly about our knowledge on the general information about the type of assessments.

These questions are mostly based on the course content, required students well-understanding of the course and requires knowledge of specific items.

The following comments, on the other hand, are on the questions written by ChatGPT.

The questions required some specific knowledge from us.

They are more flexible. Students might answer those questions with general understanding of course content and students can answer them with common sense.

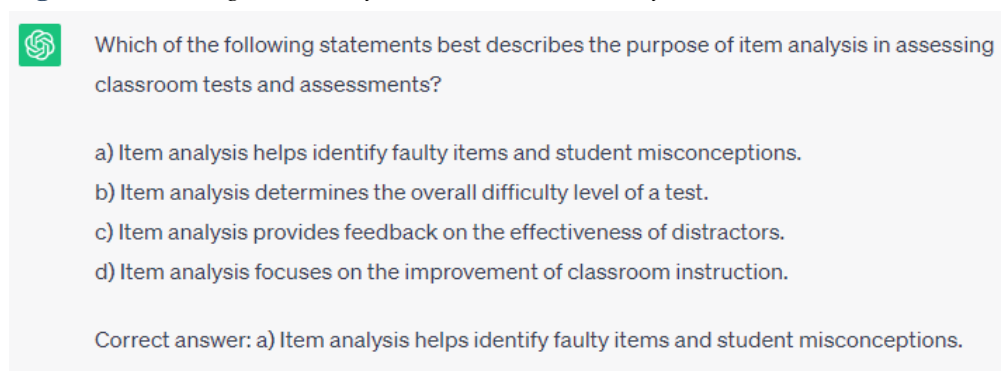
As the analysis of the students’ comments demonstrates, they had conflicting views on various aspects of the questions written by the instructor and ChatGPT.

4. DISCUSSION

The results of the study demonstrate that ChatGPT can be utilized to create classroom tests that are on par with instructors' tests as this case study shows. The chatbot was able to create multiple-choice items and when they were put together and implemented, the test produced an acceptable degree of reliability. Since the items were created according to set learning outcomes and content, supposed content validity is likely to be high. However, there were also issues with the process.

As mentioned in the methods section above, the test was created according to a table of specifications. To generate a test using ChatGPT to fit the table of specifications, I asked ChatGPT to create questions one by one by providing content. Yet, ChatGPT was not always apt to create items that were free of error. In other words, it created incorrect items. For example, when asked to create a test item about item analysis, it generated the question shown in [Figure 1](#).

Figure 1. An item generated by ChatGPT on item analysis



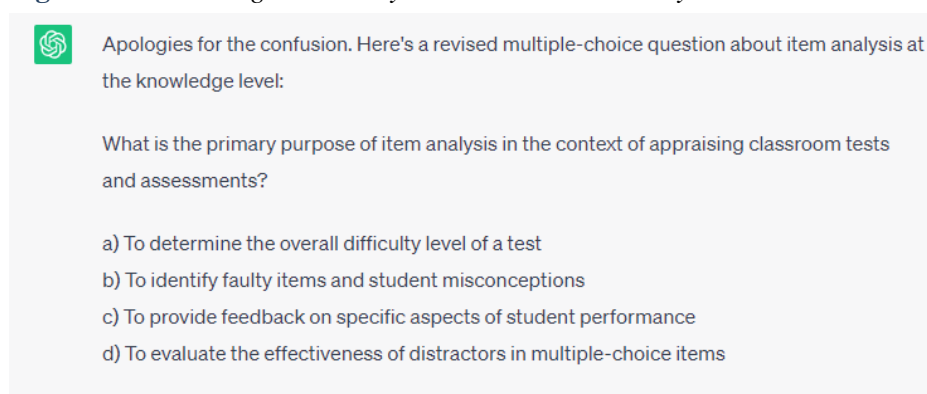
Which of the following statements best describes the purpose of item analysis in assessing classroom tests and assessments?

- a) Item analysis helps identify faulty items and student misconceptions.
- b) Item analysis determines the overall difficulty level of a test.
- c) Item analysis provides feedback on the effectiveness of distractors.
- d) Item analysis focuses on the improvement of classroom instruction.

Correct answer: a) Item analysis helps identify faulty items and student misconceptions.

As can be seen, there is not one clear correct answer to this question, although ChatGPT shows a correct answer. Other options are also among the purposes of item analysis. When asked to regenerate by highlighting the problem, ChatGPT regenerated a similar question, still with the same problem of having more than one correct answer, as shown in [Figure 2](#).

Figure 2. An item regenerated by ChatGPT on item analysis



Apologies for the confusion. Here's a revised multiple-choice question about item analysis at the knowledge level:

What is the primary purpose of item analysis in the context of appraising classroom tests and assessments?

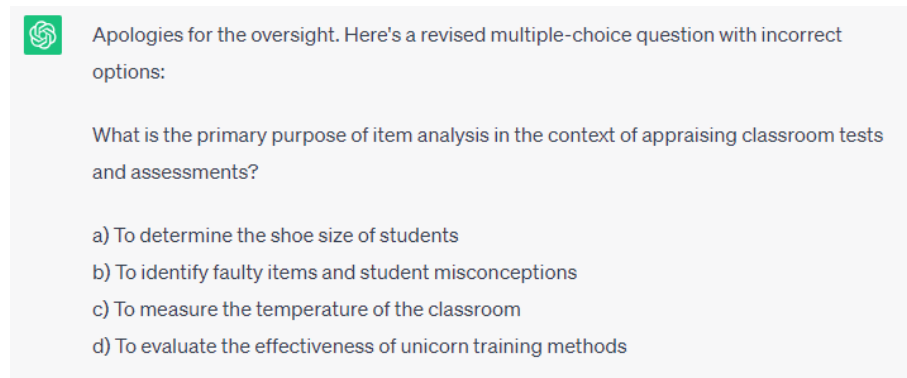
- a) To determine the overall difficulty level of a test
- b) To identify faulty items and student misconceptions
- c) To provide feedback on specific aspects of student performance
- d) To evaluate the effectiveness of distractors in multiple-choice items

When the issue of having more than one correct answer was fed into ChatGPT, it generated one with one outstanding correct answer. Yet, this time the other alternatives were nonsensical. As [Figure 3](#) shows, all three items were utterly irrelevant. It generated, for example, an alternative that read “to evaluate the effectiveness of unicorn training methods.”

This shows that ChatGPT is not always a reliable source to generate tests. It should be used with caution as it can pose a variety of failures (see, for example, Borji, 2023). Any test that ChatGPT generates should be closely scrutinized for any erroneous items. If specific items are obtained from ChatGPT, the instructor should also look for any possible errors like the ones

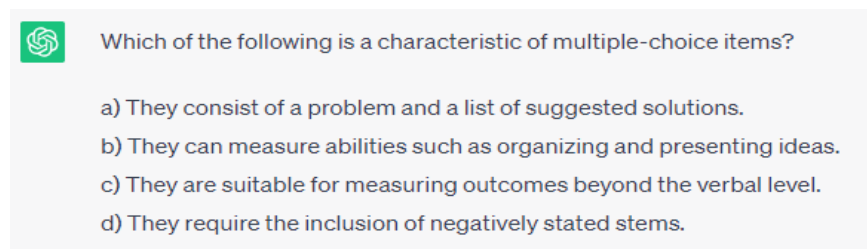
highlighted above. This leads to a legitimate question: does this take away some of the burden from the educators or rather add a more challenging task of evaluating items? This has to be carefully considered before making judgments about utilizing AI tools like ChatGPT for this purpose and in this manner.

Figure 3. An item regenerated by ChatGPT on item analysis



Another issue was to get ChatGPT to generate questions according to the learning outcomes at different levels in the cognitive domain (Anderson et al., 2001; Bloom et al., 1956). To follow the table of specifications, the prompts indicated the cognitive level the question targets. For example, ChatGPT generated the following question to a prompt where asked to create a question at the level of understanding.

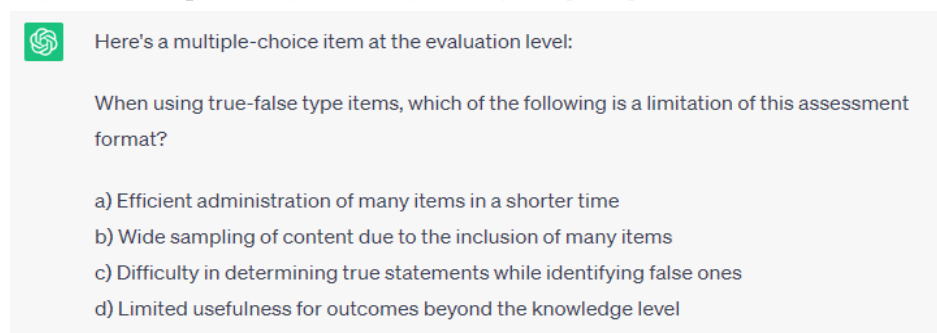
Figure 4. The question generated for the given prompt



This question (Figure 4) is not at the understanding level. One factor that differentiates items at the level of understanding from those at the level of remembering is novelty (Miller et al., 2013). ChatGPT used the wording in the content provided to it to generate the question, which means that the students can memorize the content and simply answer this question without demonstrating their understanding of the content.

The following figure (Figure 5) points to the same problem. The chatbot was asked to create a question at the level of evaluation, yet ChatGPT failed to generate one at that level as it used the same content provided to it, producing a question at the level of remembering instead.

Figure 5. The question generated for the given prompt



In a study by Moore et al. (2022), GPT-3 was used to evaluate the short-answer questions generated by the students. One thing that the researchers had GPT-3 do was to assign questions

to Bloom's taxonomy levels. Results indicate that GPT-3 failed to match the expert judgment of the cognitive level in 68% of the questions. GPT-3 also assigned 17 questions (14%) to evaluate and create levels that did not exist according to the pedagogical expert. The results of the cited study are relevant to the procedures followed in the current study where ChatGPT did not create questions at the intended level of Bloom's taxonomy. As these examples demonstrate, instructors should approach ChatGPT with caution. If the items generated by ChatGPT are used with confidence, then the tests created may not meet the need or may have low content validity, which can simply mean that they test "nothing" (Thompson & Vacha-Haase, 2018, p. 231).

In this case study, the analyses conducted have been post-hoc type such as reliability coefficient, item difficulty index, discrimination index, and distractor analyses. One potential issue with such post-hoc analyses is that the quality of the items is not tested prior to giving them to students and facing the risk of testing the students with low-quality items (Moore et al., 2023) and as such Moore et al. (2023) suggest a priori rule-based evaluation of items prior to using them for assessment. Still, both methods can be used in tandem to ensure the assessment of students' performance with the right tools and instruments. Even questions that are pre-evaluated can be analyzed through post-hoc techniques to ensure sound assessment, and it seems that both item generation and item evaluation can be handled with the assistance of AI tools such as ChatGPT before implementing specific classroom assessment tasks. Such AI tools are likely to be utilized for post-hoc analyses as well if the results are fed into them. Thus, AI tools can make assessment and evaluation tasks potentially less onerous for course instructors than they are now with the right content and prompts fed to them.

One interesting result in the reliability analysis in this case study is that the coefficient calculated for the combined test, including the instructor's and ChatGPT's items, was higher than the individual sets of tasks written by the instructor and ChatGPT alone. This finding is interesting as it may indicate that a combination of human and AI contributions may lead to an improved procedure. Halaweh (2023) asserts that "educators should encourage the use of human-AI tool augmentation for performing tasks such as finding information and ideas, editing texts and improving writing. By combining ChatGPT and human authors, the output is superior in terms of creativity, originality, and efficiency than if either one was to work alone" (p. 4).

As mentioned above, the items generated by ChatGPT were monitored by the instructor to establish that they fit with the table of specifications. Thus, the chatbot did not write a whole exam independently. This close monitoring of the questions may not reflect the independent use of AI language models to generate exams. This is relevant to González-Calatayud et al.'s (2021) contention that "this technology needs to be humanized. Research so far shows that a machine cannot assume the role of a teacher, and the way artificial intelligence works and carries out processes in the context of teaching is far from human intelligence" (p. 12). There may be ways to have AI tools to generate exams for the intended purposes of a class teacher, yet the experience in this study supports this position. The instructor needed to guide ChatGPT in preparing a test. Future research may reflect on comparing different exam generation methods like those including different degrees of contribution by the human or lack thereof.

One of the factors that are considered in addition to validity and reliability is practicality, or usability, which is related to factors such as economy, convenience, applicability, and the like (Miller et al., 2013; Thorndike & Thorndike-Christ, 2014). Since chatbots like ChatGPT or other similar AI tools can save time and effort on the part of the teachers if implemented efficiently, it would not be wrong to argue that they can increase the practicality of a test, and as such they can be said to potentially contribute to an important aspect of measurement and evaluation.

5. CONCLUSION

Research may indicate the utility of AI technologies like ChatGPT and may validate their effectiveness for use in education. However, there is also the practical aspect of the matter. Even when the use of such tools is strongly supported by empirical evidence in experimental conditions, how ready the teachers are for them is another essential issue. Wang et al. (2021) investigated, for example, the factors influencing teachers' intention to use AI technologies in teaching and found that perceived ease of use and self-efficacy were the most influential factors leading to teachers' behavioral intention to use AI technologies. They conclude that if action is taken to train teachers to enhance their self-efficacy beliefs, their attitude towards AI technologies and further intention to use them will likely increase. Thus, without incorporating such tools and their use into teacher training programs, informed practices about AI technologies to benefit teachers' experiences and students' learning will be a challenging task. Nazaretsky et al. (2022) share similar sentiments. In their research, the teachers may develop trust in AI technologies if they understand AI, AI-related technologies, and their usefulness and suggest professional development programs should include such components. In their study, teachers understood how AI works in assessing with a rubric and became more accepting of the procedures incorporating AI technologies in assessment. This indicated that to seriously consider incorporating AI tools in education, both teacher education programs and in-service training programs should be revised to include modules to prepare teachers for AI-supported practices. It is not only relevant at the individual level, organizations may also be AI-ready. Luckin et al. (2022), for example, propose a contextualized 7-step framework that will be tailored to the needs of the specific organization to help them with AI readiness.

This study aims to test the utility of ChatGPT in one aspect of the educational process in simulated testing rather than a real test situation where students would receive grades. This study was also limited to a compact group of learners enrolled in a single course at a university. More comprehensive studies eliminating such limitations are needed to further research on the issue.

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:** Final International University, Ethics Committee, 2023/019/03.

Orcid

Mehmet Kanik  <https://orcid.org/0000-0002-1737-7678>

REFERENCES

- Al-Worafi, Y.M., Hermansyah, A., Goh, K.W., Ming, L.C. (2023). Artificial intelligence use in university: Should we ban ChatGPT? preprints.org, 2023020400. <https://doi.org/10.20944/preprints202302.0400.v1>
- Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for teaching, learning, and assessment: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The Classification of Educational Goals*. David McKay.
- Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.
- Chen, G., Yang, J., Hauff, C., & Houben, G.J. (2018). LearningQ: A large-scale dataset for educational question generation. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* (Vol. 12, No. 1). Association for the Advancement of Artificial Intelligence.

- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, 20(1), 1-22.
- Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, 53, 103662.
- Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement*. Prentice-Hall.
- Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?". *Journal of Computer Assisted Learning*, 37(5), 1207-1216.
- González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12), 5467.
- Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, 15(2), ep421. <https://doi.org/10.30935/cedtech/13036>
- Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*. 13(4), 410, 1-15. <https://doi.org/10.3390/educsci13040410>
- Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence*, 3, 100076.
- McCowan, R.J., & McCowan, S.C. (1999). *Item analysis for criterion-referenced tests*. Center for Development of Human Services. <https://files.eric.ed.gov/fulltext/ED501716.pdf>
- Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. <https://doi.org/10.2139/ssrn.4354422>
- Miller, D.M., Linn, R.L., & Gronlund, N.E. (2013). *Measurement and assessment in teaching*. Pearson.
- Moore, S., Nguyen, H. A., Bier, N., Domadia, T., & Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. In I. Hilliger, P. J. Munoz-Merino, T. D. Laet, A. Ortega-Arranz & T. Farrell (Eds.), *Educating for a new future: Making sense of technology-enhanced learning adoption* (pp. 243-257). Springer.
- Moore, S., Nguyen, H.A., Chen, T., & Stamper, J. (2023). Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods. In O. Viberg, I. Jivet, P. K. Munoz-Merino, M. Perifanou & T. Papathoma (Eds.), *Responsive and Sustainable Educational Features* (pp. 229-245). Springer.
- Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British journal of educational technology*, 53(4), 914-931.
- Okonkwo, C.W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, 2, 100033. <https://doi.org/10.1016/j.caeai.2021.100033>
- Olney, A.M., Pavlik Jr, P.I., & Maass, J.K. (2017, June). Improving reading comprehension with automatically generated cloze item practice. In International Conference on Artificial Intelligence in Education (pp. 262-273). Cham: Springer International Publishing.
- Reynolds, C.R., Livingston, R.B., & Willson, V. (2009). *Measurement and assessment in education*. Pearson.
- Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J.M., Milligan, S., ... & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, 100075.
- Thompson, B., & Vacha-Haase, T. (2018). Reliability. In C. Secolsky and D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (pp. 231-251). Routledge.
- Thorndike, R.M., & Thorndike-Christ, T. (2014). *Measurement and evaluation in psychology and education*. Pearson.

-
- Van Dis, E.A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C.L. (2023). ChatGPT: five priorities for research. *Nature*, 614(7947), 224-226. <https://doi.org/10.1038/d41586-023-00288-7>
- Wang, Y., Liu, C., & Tu, Y.F. (2021). Factors affecting the adoption of AI-based applications in higher education. *Educational Technology & Society*, 24(3), 116-129.
- Yang, A.C.M., Chen, I.Y.L., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society*, 24(3), 147–158. <https://www.jstor.org/stable/27032862>