International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 622–646
https://doi.org/10.21449/ijate.1467476
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# AI-based feedback tools in education: A comprehensive bibliometric analysis study

**Mehmet Dönmez**[1*]

[1]Middle East Technical University, Ankara, Türkiye

**ARTICLE HISTORY**

*Received: Apr. 11, 2024*

*Accepted: Aug. 15, 2024*

**Keywords:**

AI-driven feedback, Educational integration, Learning enhancement, Personalized learning, Bibliometric analysis.

**Abstract:** This bibliometric analysis offers a comprehensive examination of AI-based feedback tools in education, utilizing data retrieved from the Web of Science (WoS) database. Encompassing a total of 239 articles from an expansive timeframe, spanning from inception to February 2024, this study provides a thorough overview of the evolution and current state of research in this domain. Through meticulous analysis, it tracks the growth trajectory of publications over time, revealing the increasing scholarly attention towards AI-driven feedback mechanisms in educational contexts. By describing critical thematic areas such as the role of feedback in enhancing learning outcomes, the integration of AI technologies into educational practices, and the efficacy of AI-based feedback tools in facilitating personalized learning experiences, the analysis offers valuable insights into the multifaceted nature of this field. By employing sophisticated bibliometric mapping techniques, including co-citation analysis and keyword co-occurrence analysis, the study uncovers the underlying intellectual structure of the research landscape, identifying prominent themes, influential articles, and emerging trends. Furthermore, it identifies productive authors, institutions, and countries contributing to the discourse, providing a detailed understanding of the collaborative networks and citation patterns within the community. This comprehensive synthesis of the literature serves as a valuable resource for researchers, practitioners, and policymakers alike, offering guidance on harnessing the potential of AI technologies to revolutionize teaching and learning practices in education.

## 1. INTRODUCTION

In recent years, the integration of Artificial Intelligence (AI) into various aspects of education has revolutionized teaching and learning practices. One significant area of AI application in education is developing and utilizing AI-based feedback tools (Chen, 2023). These tools, leveraging machine learning algorithms and natural language processing capabilities, offer personalized and timely feedback to students, facilitating their learning process and enhancing educational outcomes (Elmaoğlu et al., 2024; Qiao & Zhao, 2023; Su & Yang, 2023). The importance of this topic lies in its potential to reshape traditional feedback mechanisms, making them more adaptive, efficient, and effective in catering to the diverse needs of learners in contemporary educational settings.

*CONTACT: Mehmet DÖNMEZ ✉ mdonmez@metu.edu.tr Middle East Technical University, Ankara, Türkiye



**e-ISSN: 2148-7456**

As educational institutions strive to meet the evolving demands of a digital era, exploring AI-based feedback tools has gained considerable momentum in educational research. These tools encompass a wide range of applications, from automated grading systems to intelligent tutoring systems capable of providing detailed performance insights to students (Palocsay & Stevens, 2008; Roldán-Álvarez & Mesa, 2024). Consequently, a rich body of literature has emerged, documenting various aspects of AI-driven feedback tools, including their development, implementation, and impact on learning outcomes.

A review of the existing literature reveals several key themes that have surfaced in research on AI-based feedback tools in education. For instance, scholars have investigated the technical aspects of these tools, examining the algorithms and methodologies underpinning their design and functionality (Lee, 2023; Lee et al., 2023). This technical exploration is crucial for understanding AI-driven feedback systems' capabilities and limitations and optimizing their performance in educational contexts. Moreover, research in this field has also focused on the pedagogical implications of AI-based feedback tools (Conrad & Hall, 2024; Wong et al., 2023). Educators and researchers are keen to explore how these tools can be integrated into instructional practices to provide personalized guidance and support to students (Wu & Tsai, 2022). By tailoring feedback to individual learning needs and preferences, AI-driven systems have the potential to foster student engagement, motivation, and self-regulated learning (Nazari et al., 2021).

AI-based feedback tools leverage machine learning algorithms and natural language processing capabilities to offer personalized and timely feedback. These tools are used in classrooms to assist with various types of student responses, including multiple-choice questions, short answer questions, essays, and other open-ended tasks. For instance, automated writing evaluation systems provide detailed feedback on grammar, style, coherence, and content quality in student essays. Ding and Zou (2024) reviewed studies on automated writing evaluation systems, highlighting their positive impact on students' writing proficiency and the generally favorable attitudes of both learners and educators towards these tools. Besides, Shi and Aryadoust (2024) reviewed studies on automated written feedback, finding that it is predominantly studied in tertiary-level language and writing classes, with a focus on English as the target language. However, they also identified research gaps. AI-based feedback tools face challenges with more complex and open-ended tasks. Providing feedback on creative writing, complex mathematical proofs, or nuanced scientific explanations can be more difficult due to the variability and subjectivity involved in these responses. For example, while an AI tool can effectively grade multiple-choice questions or provide grammar corrections, evaluating the creativity and originality of a story or the logical coherence of a complex argument requires more sophisticated analysis that current AI technologies are still developing.

Furthermore, studies have investigated the impact of AI-based feedback tools on learning outcomes and academic achievement (Hopgood & Hirst, 2007; Téllez et al., 2024). For example, Soofi and Ahmed (2019) also systematically reviewed the studies on Intelligent Tutoring Systems and concluded that learner performance was the major method for these systems. By analyzing student performance data and feedback interactions, researchers seek to assess the effectiveness of these tools in promoting learning gains and enhancing the quality of education delivery. Understanding the causal mechanisms underlying the relationship between AI-driven feedback and learning outcomes is vital for informing evidence-based educational practices and policies (Cowling et al., 2023; Rad et al., 2023).

Despite the growing interest in AI-based feedback tools in education, there remains a need for a comprehensive bibliometric analysis to synthesize the extant literature, identify research trends, and uncover emerging themes in the field. Such an analysis holds several benefits for advancing our understanding of AI-driven feedback tools and their implications for educational practice. Mainly, a bibliometric analysis provides a systematic and objective overview of the

scholarly landscape surrounding AI-based feedback tools in education. By mapping out the volume of publications, citation networks, and collaboration patterns among researchers and institutions, this analysis offers valuable insights into the dissemination and impact of research in the field. Moreover, a bibliometric analysis facilitates the detection of research gaps and emerging trends within AI-based feedback tools in education. By analyzing keyword co-occurrence and clustering techniques, researchers can identify primary research areas and hotspots of innovation, guiding future inquiry and agenda in the field.

Based on this background, the present study aims to conduct a comprehensive bibliometric analysis of AI-based feedback tools, focusing on the domain of education and covering the publications up to February 2024. By addressing the following research questions, this study seeks to elucidate the main themes, trends, and research areas within the field:

1. What are the main themes and trends in AI-based feedback tools research within the field of education across the available literature?
2. Which countries, academic journals, and affiliations have made significant contributions to the literature on AI-driven feedback tools in education?
3. What are the primary research areas and emerging topics identified as hotspots within the field of AI-based feedback tools in education based on a comprehensive bibliometric analysis?

By undertaking this bibliometric analysis, this study tracks the trajectory of publications over time, revealing an increasing scholarly focus on AI-driven feedback mechanisms in education. Critical thematic areas explored include the role of feedback in enhancing learning outcomes, the integration of AI technologies into educational practices, and the efficacy of AI-based tools in facilitating personalized learning experiences. Through sophisticated bibliometric mapping techniques, such as co-citation and keyword co-occurrence analyses, the study uncovers the intellectual structure of the research landscape. Co-citation analysis identifies articles that are frequently cited together, highlighting seminal works and intellectual connections. On the other hand, keyword co-occurrence analysis reveals common themes and topics based on shared keywords, providing insights into prevalent research areas. These methods were chosen for their ability to systematically map the scholarly landscape, uncovering emerging trends and key contributions in the literature.

Furthermore, this study identifies key contributors (authors, institutions, and countries) engaged in advancing research in this domain, illuminating collaborative networks and citation patterns within the scholarly community. This comprehensive synthesis of the literature serves as a valuable resource for researchers, practitioners, and policymakers alike, offering strategic insights into harnessing the potential of AI technologies to revolutionize teaching and learning practices in education.

## 2. RELATED WORK

Artificial Intelligence (AI) has influenced various domains from revolutionizing processes to practices, and including education. In recent years, the integration of AI into educational settings has garnered significant attention, with researchers and educators exploring its potential to enhance teaching and learning outcomes (Kim & Adlof, 2024; Li et al., 2024). One featured area of AI application in education is the development and utilization of AI-based feedback tools. These tools leverage advanced algorithms and natural language processing capabilities to provide personalized and timely feedback to learners, aiming to improve their performance and engagement in educational activities (Farshad et al., 2023; Fu et al., 2020; Kumar & Boulanger, 2020).

The integration of AI-driven feedback tools into education is motivated by several factors. Firstly, traditional feedback methods, such as manual grading and assessments, are often time-consuming and resource-intensive for educators (Gao et al., 2024). With growing class sizes

and diverse learner needs, there is a pressing need for scalable and efficient feedback mechanisms to accommodate modern education systems' demands. AI-based feedback tools offer a promising solution by automating the feedback process, thereby freeing up educators' time and resources to focus on more value-added tasks (Zhao et al., 2023). For instance, AI-powered grading systems can quickly evaluate and score large volumes of student essays, providing detailed feedback on writing quality, grammar, and coherence, which can be particularly useful in writing-intensive courses (Yavuz et al., 2024).

Moreover, AI-driven feedback tools have the potential to address the challenge of personalized learning in education. Every learner has unique strengths, weaknesses, and learning preferences, necessitating tailored instructional strategies and feedback mechanisms (Kubsch et al., 2022). However, providing individualized feedback to each student in a traditional classroom setting can be challenging due to time constraints and logistical limitations. AI-based feedback tools overcome this challenge by analyzing vast amounts of student data and generating personalized feedback that is tailored to each learner's needs, including those of children with special needs (Ebenbeck & Gebhardt, 2024). For example, adaptive learning platforms can use AI to assess student performance in real-time and provide customized learning paths and resources, ensuring that each student receives the appropriate level of challenge and support (Gligorea et al., 2023).

Furthermore, AI-driven feedback tools hold promise for promoting self-regulated learning and metacognitive skills development among students (Hopfenbeck et al., 2023; Liang et al., 2024). Research has shown that effective feedback is crucial in facilitating students' ability to monitor and regulate their own learning processes (Zheng et al., 2021). By providing timely and actionable feedback, AI-driven tools empower students to reflect on their performance, identify areas for improvement, and take proactive steps to enhance their learning outcomes (Sharma et al., 2019). For instance, AI-based systems can track student progress over time and provide insights into study habits and learning strategies, encouraging students to develop better self-assessment and planning skills (Li & Kim, 2024). Thus, integrating AI-based feedback tools into educational settings has the potential to foster a culture of continuous improvement and self-directed learning among students.

Despite the potential benefits of AI-based feedback tools, their integration into educational practice is not without challenges. One key challenge is ensuring the validity and reliability of the feedback generated by these tools (Kaldaras et al., 2022). As AI algorithms rely on statistical models and machine learning techniques, there is a risk of bias or error in the feedback provided. Educators and researchers must critically evaluate the accuracy and appropriateness of AI-generated feedback to ensure its utility and effectiveness in supporting student learning (Wang et al., 2024). An example of this issue is the need to regularly update and validate the algorithms used in automated essay scoring to avoid perpetuating any biases present in the training data (Bui & Barrot, 2024).

Additionally, the ethical implications of AI-driven feedback tools require careful consideration (Su & Yang, 2023; Wong et al., 2023). These tools often involve the collection and analysis of sensitive student data, raising concerns about privacy, security, and data protection (Chavez et al., 2023; Williams, 2024). Educators and policymakers must navigate these ethical dilemmas and establish robust safeguards to protect students' rights and interests while using AI's potential in education. For example, implementing strict data anonymization protocols and transparency measures can help decrease privacy risks associated with AI-driven systems (Shahriar et al., 2023).

In the field of research, there has been a growing interest in exploring the design, implementation, and impact of AI-based feedback tools in education. Most of the studies have investigated various aspects of these tools, including their technical underpinnings, pedagogical implications, and effects on student learning outcomes. For example, researchers have

developed AI-driven feedback systems for automated grading and assessment, personalized tutoring, and formative feedback provision (Palocsay & Stevens, 2008; Roldán-Álvarez & Mesa, 2024). These studies have yielded valuable insights into AI-driven feedback tools' potential applications and limitations in educational contexts.

Moreover, scholars have examined the factors influencing the adoption and acceptance of AI-based feedback tools among educators and students (Chiu et al., 2022). Understanding their perceptions, attitudes, and experiences is essential for informing the design and implementation of effective feedback systems. Additionally, research has explored the role of AI-driven feedback in promoting equity and inclusivity in education by addressing disparities in access to personalized support and resources among diverse learner populations (Khoo & Kang, 2022). For instance, AI tools can be used to identify and support at-risk students by providing early intervention strategies tailored to their specific needs (Nimy et al., 2023).

Overall, the literature on AI-based feedback tools in education is massive and complicated, reflecting the diverse interests and perspectives of researchers and practitioners. However, despite the wealth of research available, there remains a need for a comprehensive bibliometric analysis to synthesize the existing literature, identify research trends, and uncover emerging themes. Such an analysis would provide valuable insights into the current state of research on AI-driven feedback tools in education and inform future directions for inquiry and innovation in the field.

## 3. METHODOLOGY

### 3.1. Inquiry Process

The study started with a bibliometric analysis to summarize prior studies using AI-based feedback tools to enhance learning experiences. A comprehensive exploration of literature concerning the utilization of AI-based feedback tools to improve learning experiences was conducted by searching the widely recognized electronic database, Web of Science (WoS). This inquiry specifically targeted educational research. On February 22, 2024, the literature within WoS was examined by using the following search string: (feedback AND (educa* OR learn* OR teach*) AND (AI OR artificial intelligence OR chatgpt)).

### 3.2. Selection Process

While selecting relevant papers, the criteria for inclusion and exclusion (as outlined in Table 1) were defined by following the PRISMA guideline for systematic literature reviews, as proposed by Page et al. (2021). Subsequently, a meticulous selection process was carried out in four distinct stages: identification, screening, eligibility assessment, and final inclusion. This systematic approach ensured a comprehensive and rigorous selection of papers that met the research objectives.

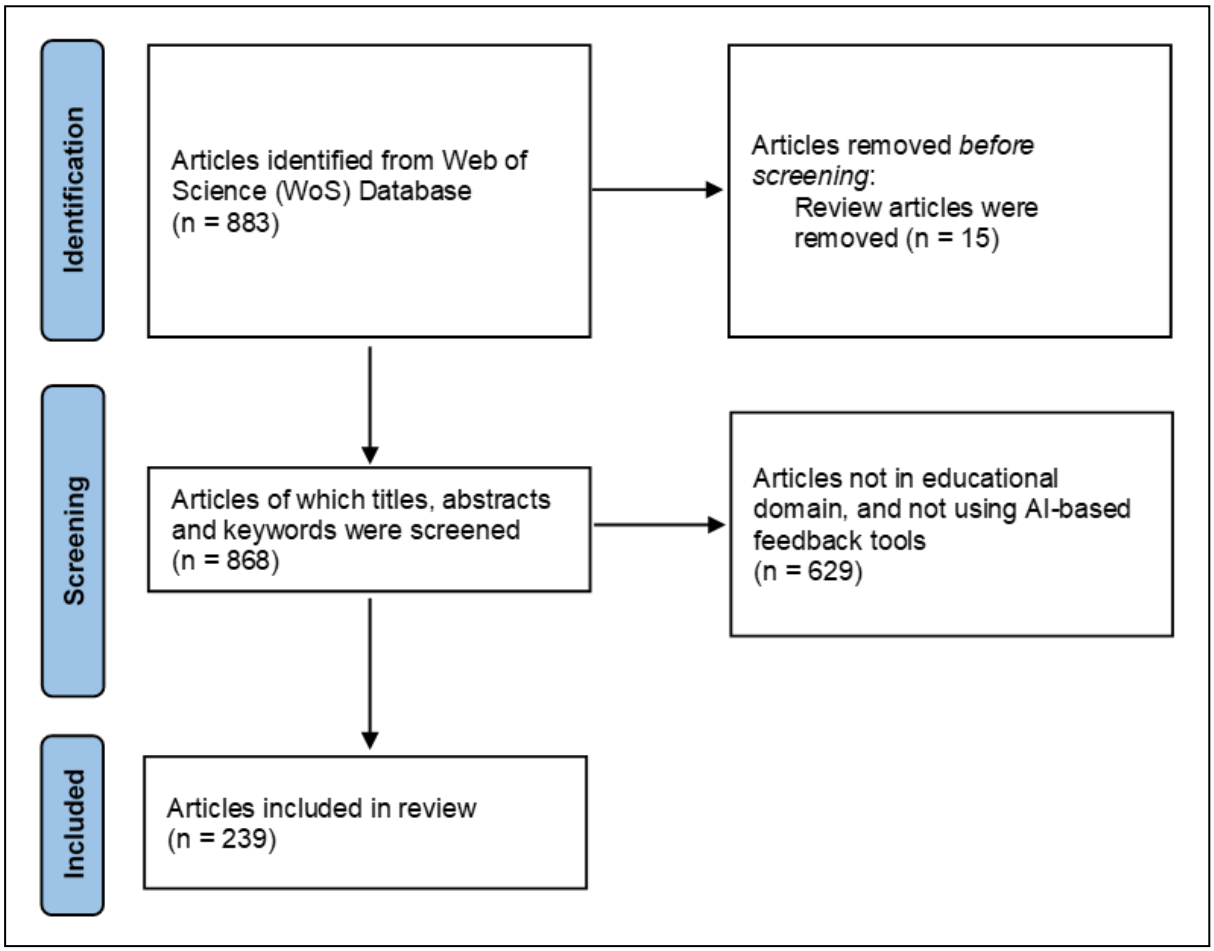
**Table 1.** *Inclusion and exclusion criteria.*

| Inclusion Criteria | Exclusion Criteria |
|---|---|
| Published in an academic journal | Review, meta-analysis, or conference paper |
| Written in English | Not written in English |
| Available in full-text | Not available in full-text |
| Research paper in the educational domain | Research paper not in the educational domain |
| Using AI-based feedback tools | Not using AI-based feedback tools |

Initially, the review of studies across the WoS database strictly followed predefined inclusion and exclusion criteria, as outlined in Table 1. A total of 883 articles were initially retrieved, from which 15 review articles were identified and removed during the initial screening phase. Following this, the titles, abstracts, and keywords of the remaining 868 articles underwent meticulous inspection to identify those aligning with the inclusion and exclusion criteria.

Consequently, an additional 629 articles not in the educational domain and not using AI-based feedback tools were excluded from consideration in this study. As a result, 239 articles were considered appropriate for inclusion in the current study. A visual representation of the inquiry and selection processes is provided in Figure 1.

**Figure 1.** *Inquiry and selection process.*




### 3.3. Data Analysis

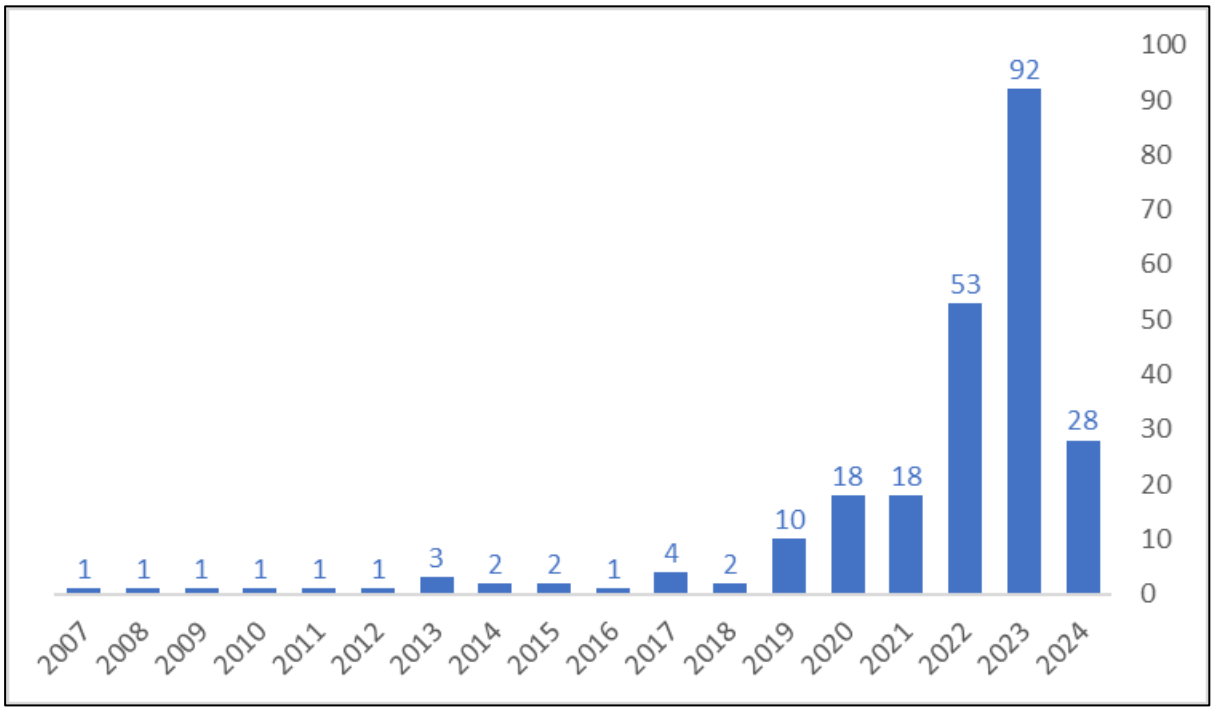
For this study, a comprehensive data analysis was conducted utilizing the WoS database. Initially, a BibTeX file was generated to encompass all pertinent data. Subsequently, the biblioshiny web interface, integrated within RStudio along with the bibliometrix package, facilitated the bibliometric analysis and visualization process (Aria & Cuccurullo, 2017). This approach provided a user-friendly interface, enabling the creation of diverse visual representations, including tables and graphs.

## 4. RESULTS and DISCUSSION

### 4.1. Descriptive Analysis

Spanning from 2007 to 2024, the analysis encompassed data extracted from 147 distinct journals among 239 publications. Figure 2 shows a significant increase in the number of publications over the years, particularly from 2019 onwards. The most notable surge occurred between 2022 and 2023, reflecting a growing interest and scholarly attention towards AI-driven feedback mechanisms in educational contexts.

**Figure 2.** *Number of publications over the years.*




The provided summary table (Table 2) offers a comprehensive overview of the bibliometric analysis conducted on AI-based feedback tools in education. Notably, the annual growth rate of the field stands at an impressive 21.65%, indicative of the increasing interest and scholarly activity surrounding AI-driven feedback mechanisms in educational settings (Kartal & Yeşilyurt, 2024; Song & Wang, 2020).

**Table 2.** *Summary of bibliometric analysis results on AI-based feedback tools.*

| Description | Results |
|---|---|
| Main Information About Data | |
| Timespan | 2007:2024 |
| Sources (Journals) | 147 |
| Documents | 239 |
| Annual Growth Rate % | 21.65 |
| Document Average Age | 2.43 |
| Average citations per doc | 7.577 |
| References | 10306 |
| Document Contents | |
| Keywords Plus (ID) | 343 |
| Author's Keywords (DE) | 829 |
| Authors | |
| Authors | 770 |
| Authors of single-authored docs | 28 |
| Authors Collaboration | |
| Single-authored docs | 28 |
| Co-Authors per Doc | 3.67 |
| International co-authorships % | 25.52 |

Exploring deeper into the document characteristics, the average age of the included documents is relatively low at 2.43 years, underscoring the currency and relevance of the literature

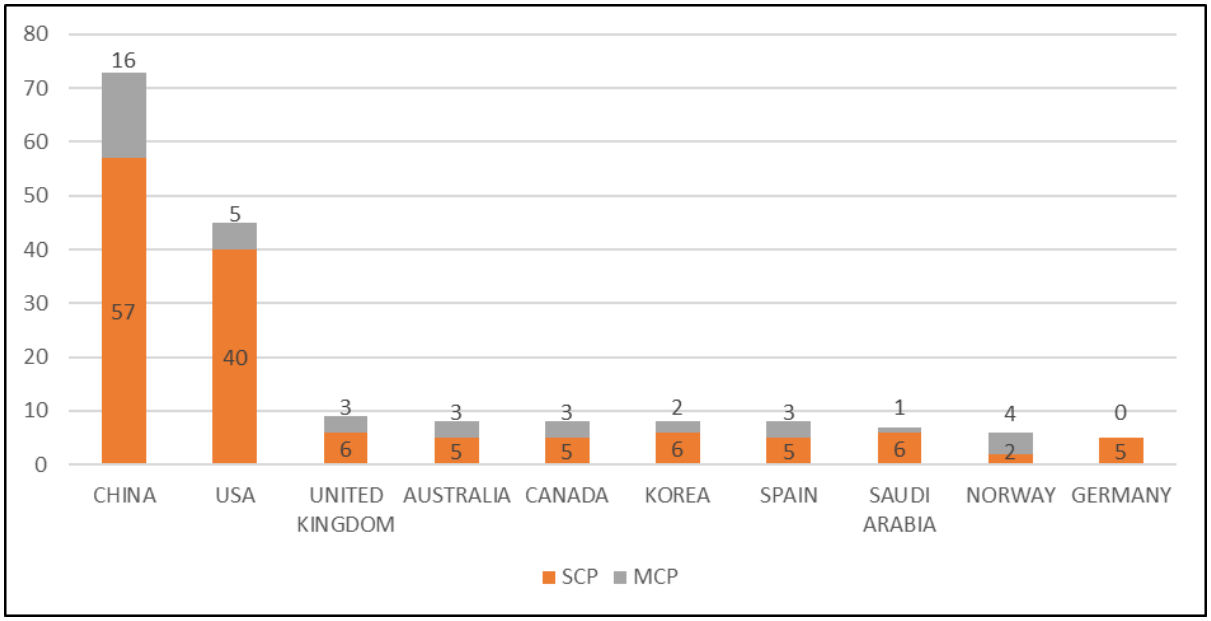
examined. Moreover, each document garners an average of 7.58 citations, indicative of the scholarly impact and influence wielded by research on AI-based feedback tools in education (Bin-Hady et al., 2023). In terms of document contents, a rich tapestry of keywords emerges, with 343 Keywords Plus and 829 author's keywords encapsulating the diverse facets and dimensions explored within the field. This range of keywords reflects the multifaceted nature of research endeavors surrounding AI-based feedback tools, encompassing technical, pedagogical, and evaluative perspectives (Rubio-Manzano et al., 2019). The analysis also sheds light on the collaborative nature of research in this domain, with 770 distinct authors contributing to the body of literature examined. Interestingly, while the majority of documents are co-authored, a notable proportion, 28 documents, are single-authored, indicative of the diverse scholarly contributions within the field. Furthermore, the collaborative landscape extends beyond national borders, with international co-authorships accounting for 25.52% of the total collaborations. This global dimension underscores the transnational collaboration and exchange of ideas characterizing research endeavors in AI-based feedback tools in education (Chen et al., 2023).

In summation, the descriptive analysis of the results provides a nuanced understanding of the breadth, depth, and collaborative dynamics inherent within the scholarly discourse surrounding AI-based feedback tools in education.

### 4.1.1. *Influential countries*

Figure 3 presents an analysis of the top 10 countries based on the corresponding authors of articles related to AI-based feedback tools in education. The data is segmented into several categories, including the number of articles authored by individuals from each country, the count of single-country publications (SCP), the count of multiple-country publications (MCP), the frequency of each country's appearance, and the ratio of multiple-country publications to total publications.

**Figure 3.** *Top 10 countries of corresponding authors.*




China emerges as the leading contributor, with 73 articles authored by corresponding authors based in the country. Among these articles, 57 are single-country publications, indicating a significant level of independent research output. However, China also demonstrates substantial collaboration with other countries, as evidenced by 16 multiple-country publications. The United States follows closely behind, with 45 articles attributed to corresponding authors from the country. Of these, 40 are single-country publications showcasing a strong domestic research

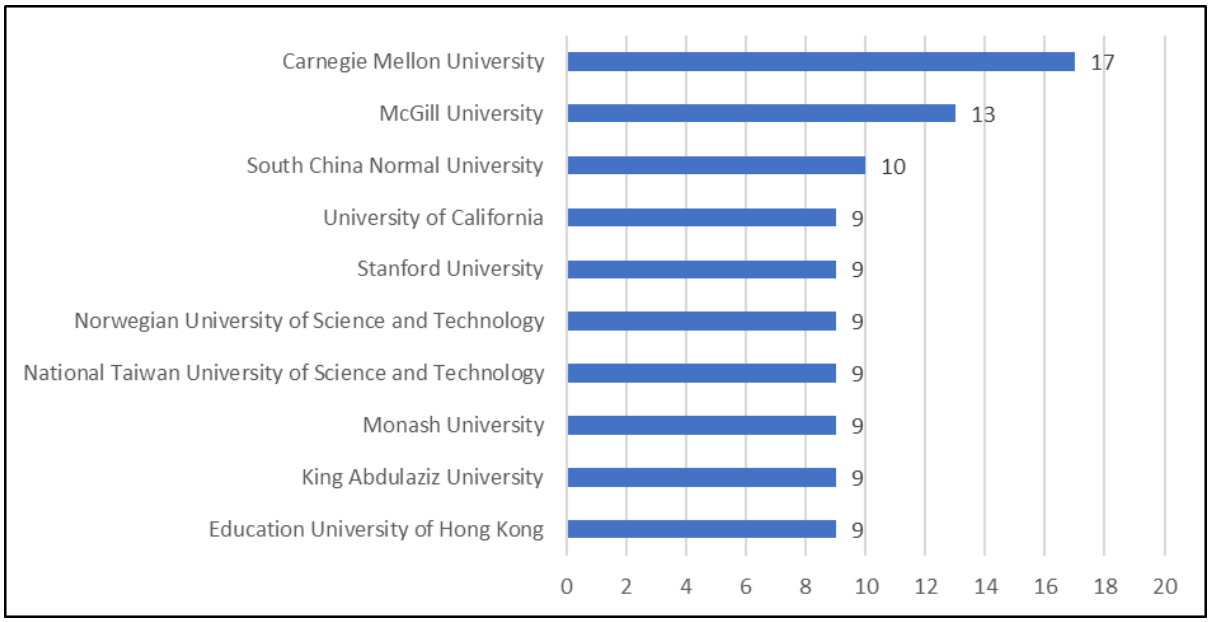
presence. The USA also engages in collaborative efforts with five multiple-country publications. Other notable contributors include the United Kingdom, Australia, Canada, Korea, Spain, Saudi Arabia, Norway, and Germany. Each of these countries has varying levels of research output and collaboration patterns. For instance, Norway stands out with a high MCP Ratio of 0.667, indicating a significant propensity for international collaboration, despite a smaller overall number of articles.

To sum up, this figure underscores the global nature of research on AI-based feedback tools in education, with contributions from diverse geographical locations. It also highlights the prevalence of both independent and collaborative research efforts, providing valuable insights into the international landscape of scholarly inquiry in this field (Zhang et al., 2024).

### 4.1.2. *Influential affiliations*

Figure 4 presents an analysis of the top 10 affiliations of corresponding authors for articles related to AI-based feedback tools in education. Each affiliation is accompanied by the number of articles attributed to corresponding authors associated with that institution.

**Figure 4.** *Top 10 affiliations of corresponding authors.*




Carnegie Mellon University emerges as the leading affiliation, with 17 articles authored by corresponding authors affiliated with the institution. It indicates a significant research presence and activity in the field of AI-based feedback tools within the Carnegie Mellon University community. Following closely behind are McGill University and South China Normal University, each with 13 and 10 articles, respectively. These affiliations also demonstrate notable research output and engagement with the topic under investigation. The list of top affiliations also includes institutions such as Education University of Hong Kong, King Abdulaziz University, Monash University, National Taiwan University of Science and Technology, Norwegian University of Science and Technology, Stanford University, and University of California, Irvine. Each of these institutions has contributed a substantial number of articles, showcasing their involvement in research related to AI-based feedback tools in education.

Overall, this figure provides valuable insights into the institutional landscape of scholarly inquiry in this field, highlighting key contributors and hubs of research activity. These affiliations play a crucial role in shaping the discourse and advancement of knowledge in AI-based feedback tools in education.

### 4.1.3. *Influential journals*

Table 3 provides an overview of the top 10 influential journals within the realm of AI-based feedback tools in education and the number of articles published in each journal.

**Table 3.** *Top 10 influential journals.*

| Journals | # of Articles |
|---|---|
| International Journal of Artificial Intelligence in Education | 14 |
| Education and Information Technologies | 9 |
| British Journal of Educational Technology | 8 |
| Sustainability | 6 |
| Applied Sciences-Basel | 5 |
| Frontiers in Education | 5 |
| Frontiers in Psychology | 5 |
| Interactive Learning Environments | 5 |
| Computers & Education | 4 |
| IEEE Transactions on Learning Technologies | 4 |

“International Journal of Artificial Intelligence in Education” is at the top of the list with 14 articles. This journal can be seen as a featured platform for scholarly discourse and research dissemination about the intersection of artificial intelligence and education, particularly focusing on feedback mechanisms. Following closely behind is “Education and Information Technologies”, with 9 articles. This journal encompasses a broad spectrum of topics related to educational technology, including the development and application of AI-based feedback tools in educational settings. The “British Journal of Educational Technology” also features prominently on the list, with 8 articles. This journal is renowned for its contributions to the field of educational technology, showcasing research on innovative methodologies and technologies, including AI-driven feedback mechanisms. Other notable journals include “Sustainability” (6 articles), “Applied Sciences-Basel” (5 articles), “Frontiers in Education” (5 articles), “Frontiers in Psychology” (5 articles), “Interactive Learning Environments” (5 articles), “Computers & Education” (4 articles), and “IEEE Transactions on Learning Technologies” (4 articles). Each of these journals plays a significant role in disseminating research findings and fostering scholarly discourse on AI-based feedback tools and their impact on educational outcomes.

Overall, the table provides valuable insights into the scholarly landscape of AI-based feedback tools in education, highlighting key journals that serve as platforms for research dissemination and knowledge exchange in this burgeoning field.

### 4.1.4. *Influential publications*

Table 4 showcases the top 10 most cited publications related to AI-based feedback tools in education, along with the authors, publication sources, purposes, and the number of citations recorded on the Web of Science (WoS) platform. The publication titled "Automated Writing Assessment in the Classroom" by Warschauer and Grimes (2008) is at the top of the list and published in Pedagogies, which has gathered 105 citations in WoS. This influential work explores the application of an automated essay assessment tool in secondary schools, utilizing interviews, surveys, and classroom observations to assess its effectiveness as a teaching tool and its influence on teachers' instructional practices and students' writing behaviors. Following closely behind is "The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine" by Mirchi et al. (2020), published in Plos One, with 92 citations. This study introduces and validates a new framework utilizing explainable artificial intelligence for simulation-based training in surgery, concluding in the development of an automated educational feedback platform, with the aim of enhancing

surgical education by providing participants with immediate, objective feedback based on proficiency benchmarks and expert classification.

**Table 4.** *Top 10 most cited publications.*

| Authors and Year | Source | Purpose | Citations on WoS |
|---|---|---|---|
| Warschauer and Grimes (2008) | Pedagogies | To investigate the implementation and impact of automated essay-scoring software in secondary school classrooms | 105 |
| Mirchi et al. (2020) | Plos One | To introduce and validate an automated educational feedback platform designed for simulation-based training in surgery and medicine | 92 |
| McLaren et al. (2011) | Computers & Education | To investigate whether employing polite feedback and hints in web-based intelligent tutoring systems impacts student learning outcomes positively | 56 |
| Cukurova et al. (2019) | British Journal of Educational Technology | To explore the potential role of artificial intelligence in education as a tool for augmenting human intelligence | 51 |
| Chin et al. (2010) | Educational Technology Research and Development | To investigate the effectiveness of Teachable Agents (TA) in K-12 education | 51 |
| Rahman and Watanobe (2023) | Applied Sciences-Basel | To investigate the potential impact of ChatGPT on education and research | 48 |
| Sharma et al. (2019) | British Journal of Educational Technology | To explore the development of pipelines for educational data leveraging artificial intelligence and multimodal analytics | 48 |
| Rose et al. (2019) | British Journal of Educational Technology | To encourage the development of explanatory learner models in education | 44 |
| Nazari et al. (2021) | Heliyon | To investigate the effectiveness of an Artificial Intelligence (AI) powered writing tool | 37 |
| Bañeres et al. (2020) | Applied Sciences-Basel | To develop and evaluate an accurate predictive model and an early warning system to identify at-risk students | 34 |

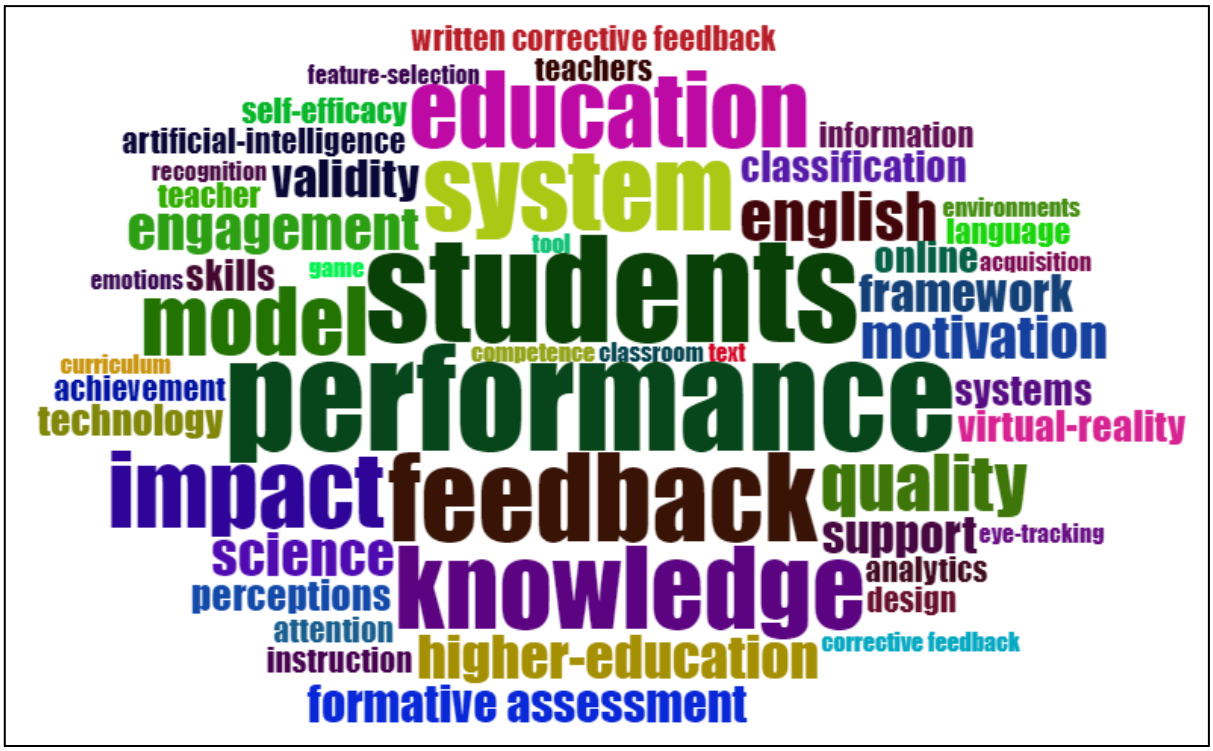
Other notable publications include "Polite web-based intelligent tutors: Can they improve learning in classrooms?" by McLaren et al. (2011) in Computers & Education (56 citations), and "Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring" by Cukurova et al. (2019) in the British Journal of Educational Technology (51 citations). Additionally, "Preparing students for future learning with Teachable Agents" by Chin et al. (2010) in Educational Technology Research and Development (51 citations), and "ChatGPT for Education and Research: Opportunities, Threats, and Strategies" by Rahman and Watanobe (2023) in Applied Sciences-Basel (48 citations), also feature prominently in the list, underscoring their impact on the discourse surrounding AI-driven educational technologies. Furthermore, "Building pipelines for educational data using AI and multimodal analytics: A 'grey-box' approach" by Sharma et al. (2019) in the British Journal of

Educational Technology (48 citations), and "Explanatory learner models: Why machine learning (alone) is not the answer" by Rosé et al. (2019) in the same journal (44 citations), highlight the importance of interpretability and transparency in AI-driven educational systems. Rounding off the list are "Application of Artificial Intelligence powered digital writing assistant in higher education: randomized controlled trial" by Nazari et al. (2021) in Heliyon (37 citations), and "An Early Warning System to Detect At-Risk Students in Online Higher Education" by Bañeres et al. (2020) in Applied Sciences-Basel (34 citations), shedding light on the diverse applications and implications of AI-based feedback tools in educational contexts.

### 4.2. Keyword Analysis

The word cloud included 50 frequent keywords and was generated from the Keywords Plus data, which highlights the prominent themes and concepts widespread in the literature related to AI-based feedback tools in education (see Figure 5).

**Figure 5.** *Most frequent 50 keywords generated from Keywords Plus.*

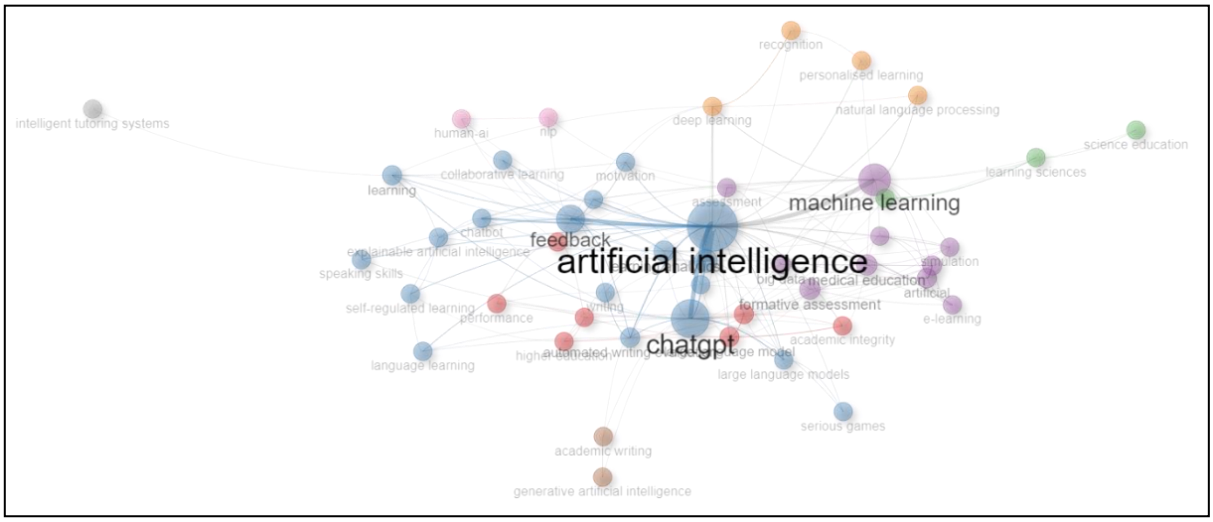



The most frequently occurring terms, "performance" and "students," underscore the central focus on student outcomes and achievement within educational contexts. These terms suggest a keen interest in assessing and enhancing student performance through the utilization of AI-driven feedback mechanisms (Afzaal et al., 2024). Additionally, the term "feedback", one of the keywords among the search terms of this study, emerges prominently, reflecting the pivotal role of feedback provision in the educational process. This emphasis on feedback aligns with the overarching goal of leveraging AI technology to deliver personalized and timely feedback to students, thereby facilitating their learning and skill development (Heeg & Avraamidou, 2023). Other notable terms include "knowledge," "system," and "education," which highlight the broader context of educational technology and the integration of AI-based systems into educational settings. These terms indicate a multifaceted approach to utilizing AI technology to enhance knowledge acquisition and educational practices (Stojanov, 2023). Furthermore, terms such as "impact" and "quality" suggest a focus on assessing the effectiveness and efficacy of AI-based feedback tools in driving positive educational outcomes. It reflects a critical examination of the impact of technology on teaching and learning processes, with an emphasis on ensuring the quality and integrity of educational interventions (Lee et al., 2024). Finally, the

inclusion of terms like "model" and "English" hints at the diversity of research interests within the field, encompassing topics such as AI modeling techniques and the application of feedback tools in specific educational domains, such as language learning (Kartal & Yeşilyurt, 2024; Shi & Aryadoust, 2024). Overall, the word cloud provides a visually compelling representation of the key themes and concepts underlying research on AI-based feedback tools in education, offering valuable insights into the prevailing trends and interests within the field.

Moreover, the thematic map depicts the author's keywords' distribution, including 50 keywords, in AI-based feedback tools in education, organized into distinct clusters based on their semantic similarities and thematic relevance (see Figure 6).

**Figure 6.** *Thematic map of author's keywords.*




Cluster 1, labeled "Artificial Intelligence (AI)," encompasses keywords related to artificial intelligence technologies, including "artificial intelligence," "large language model," "higher education," "research," "performance," and "students." These keywords reflect the overarching focus on AI-driven approaches to feedback provision and educational enhancement. For instance, recent studies by Ouyang et al. (2023) and Rad et al. (2023) highlight how AI, particularly large language models, improves feedback quality and student engagement in higher education contexts.

Cluster 2, also under the label "Artificial Intelligence (AI)," predominantly features keywords associated with specific AI applications in education, such as "ChatGPT," "learning analytics," "automated writing evaluation," and "personalized feedback." This cluster highlights the diverse range of AI-based tools and methodologies utilized for educational purposes, including chatbots, analytics platforms, and automated assessment systems (Chang et al., 2023; Ding & Zou, 2024).

Cluster 3, labeled "Educational Technology," encompasses keywords related to educational technology and instructional design, such as "educational technology," "learning sciences," and "science education." These keywords underscore the intersection between AI-driven feedback tools and broader educational technology frameworks, emphasizing the integration of technology into pedagogical practices (Sağın et al., 2023).

Cluster 4, labeled "Machine Learning," comprises keywords related to machine learning algorithms and methodologies, including "machine learning," "assessment," "formative assessment," and "big data." This cluster highlights the increasing adoption of machine-learning techniques for analyzing educational data, providing personalized feedback, and optimizing instructional strategies (Jaleniauskienė et al., 2023).

Cluster 5, labeled "Deep Learning," focuses on keywords associated with deep learning techniques, such as "deep learning," "natural language processing," and "recognition." These

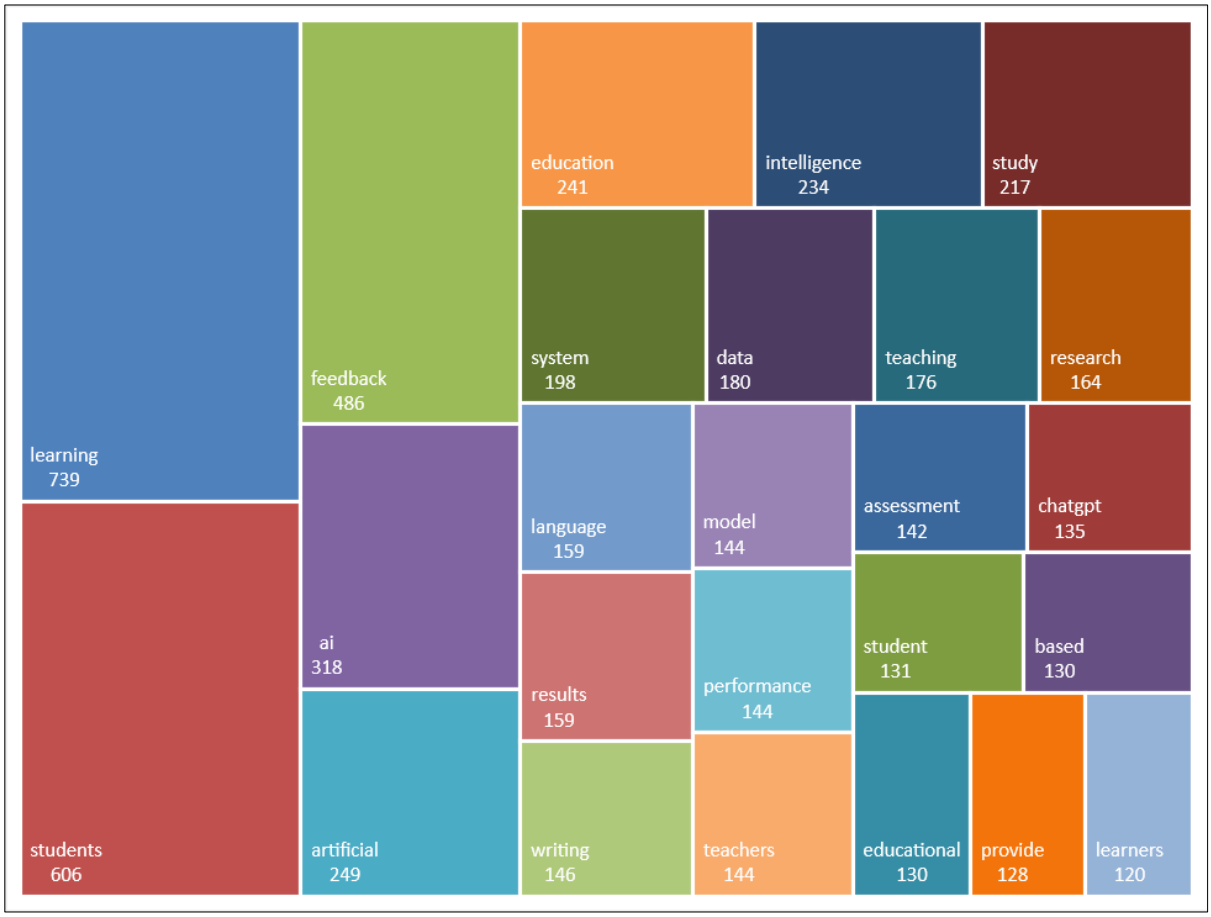
keywords signify the growing interest in deep learning models for processing and analyzing educational data, particularly in the context of natural language understanding and recognition tasks (Li & Mohamad, 2023).

Cluster 6, labeled "Academic Writing," includes keywords related to academic writing and text generation, such as "academic writing" and "generative artificial intelligence." This cluster suggests a specific focus on AI applications in academic writing support and text generation tools (Barrett & Pack, 2023).

Cluster 7, labeled "Human-AI Interaction," encompasses keywords related to the interaction between humans and AI systems, including "human-AI" and "NLP" (Natural Language Processing). This cluster highlights the importance of considering human factors and user experiences in the design and implementation of AI-based feedback tools in education (Wang et al., 2024).

Lastly, Cluster 8, labeled "Intelligent Tutoring Systems," features keywords related to intelligent tutoring systems, such as "intelligent tutoring systems." This cluster focuses on AI-driven tutoring systems that provide students with personalized learning experiences and adaptive feedback. Gu (2024) and Roldán-Álvarez and Mesa (2024) highlight how intelligent tutoring systems leverage AI technologies to tailor educational content and feedback to individual student needs, thereby improving learning outcomes.

**Figure 7.** *Most frequently used 25 words in abstracts.*



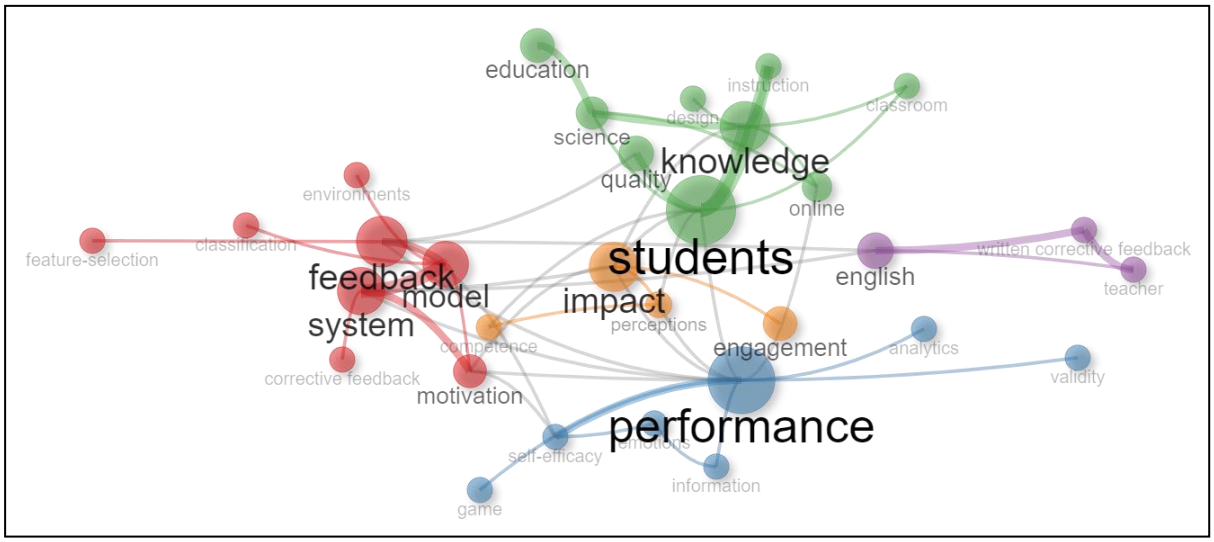

Overall, the thematic map provides a comprehensive overview of the key themes and topics within the field of AI-based feedback tools in education, highlighting the diverse range of AI applications, educational technologies, and pedagogical approaches utilized in research and practice.

Furthermore, Figure 7 presents a visualization of the most frequent 25 words extracted from abstracts of scholarly articles on AI-based feedback tools in education. At the forefront of these words is "learning," indicating a primary focus on educational processes and outcomes within the literature. Subsequently, "students" and "feedback" emerge as prominent themes, underscoring the importance of student engagement and assessment in the context of AI-driven educational interventions. The terms "AI" and "artificial intelligence" reflect the pervasive use of AI technologies in educational settings, particularly in feedback provision and personalized learning experiences. Moreover, key concepts such as "education," "teaching," "research," and "assessment" highlight the multifaceted nature of research endeavors in this field, encompassing pedagogical practices, empirical investigations, and evaluative methodologies. Additionally, the presence of specific terms like "language," "writing," and "ChatGPT" suggests a focus on language learning, writing instruction, and the integration of AI-powered chatbots in educational environments. Overall, the figure provides a brief overview of the prevalent themes and topics addressed in the abstracts of scholarly articles related to AI-based feedback tools in education, offering insights into the scope and depth of research conducted in this domain.

### 4.3. Conceptual Analysis

The co-occurrence network analysis based on Keywords Plus was utilized to reveal potential research topics along with their relationships and to interpret the knowledge embedded within thematic clusters in the field of AI-based feedback tools in education, providing insights into the relationships between different concepts. The default parameters of the "bibliometrix" package on the web interface "biblioshiny" were employed, including the utilization of the "Walktrap" clustering algorithm with 50 keywords and a minimum of two edges. The obtained five clusters from 31 nodes are depicted in Figure 8.

**Figure 8.** *Co-occurrence network based on Keywords Plus.*

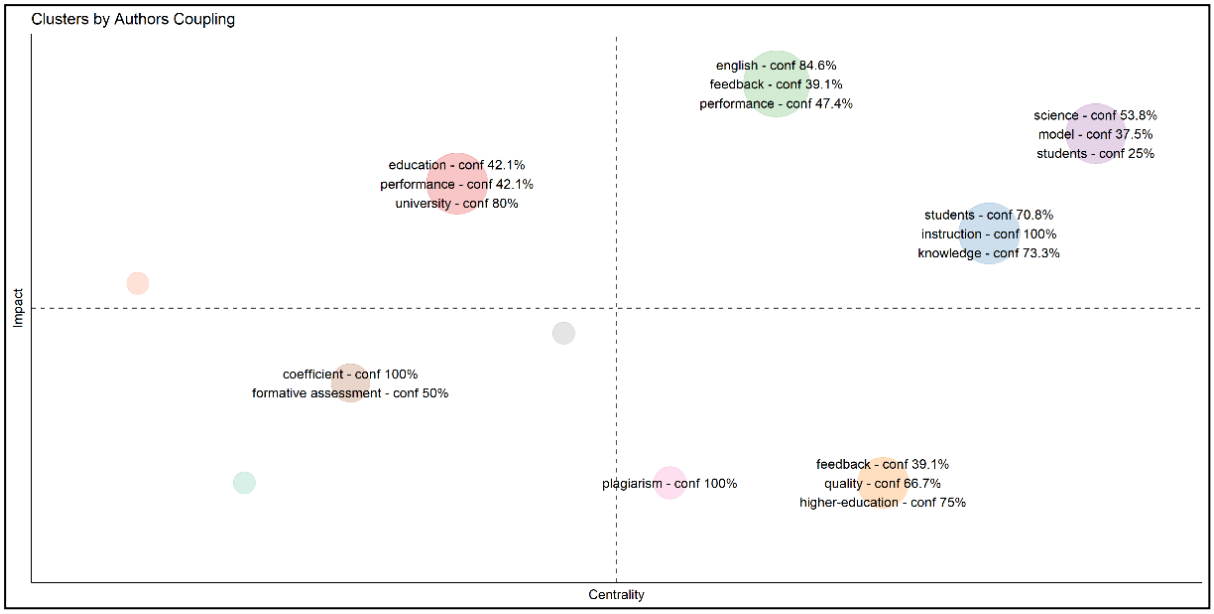



In Cluster 1, terms such as "feedback," "system," "model," and "motivation" emerge as central nodes with high betweenness, closeness, and PageRank centrality scores. These terms represent fundamental components of feedback systems in educational settings, highlighting their significance in research and practice. Cluster 2 focuses on terms related to "performance," "validity," "analytics," and "self-efficacy," indicating a strong emphasis on assessing and optimizing learning outcomes through AI-driven feedback mechanisms. These terms suggest a particular interest in leveraging data analytics and machine learning techniques to enhance performance evaluation and learner motivation. Cluster 3 encompasses terms like "students," "knowledge," "education," and "quality," underscoring the importance of student-centered approaches to education and the pursuit of high-quality learning experiences. These terms

reflect a holistic view of education, emphasizing the acquisition of knowledge and the promotion of educational excellence. In Cluster 4, terms such as "English," "teacher," and "written corrective feedback" suggest a focus on language learning and pedagogical practices in the context of AI-based feedback tools. These terms highlight the role of technology in supporting language instruction and providing personalized feedback to learners. Cluster 5 includes terms like "impact," "engagement," and "perceptions," indicating an interest in understanding the effects of AI-based feedback tools on student engagement and perceptions of learning. These terms suggest a broader consideration of the socio-emotional aspects of education and the implications of technology integration on student outcomes. Overall, the co-occurrence network offers a comprehensive view of the interconnected nature of key concepts in AI-based feedback tools research, illustrating the multidimensional relationships between different aspects of educational practice and technology utilization.

Moreover, Figure 9 illustrates clustering by coupling among the authors measured by Keyword Plus, with cluster labeling also based on Keyword Plus and impact measured by global citation score. The following parameters were utilized: (i) restricting the analysis to 250 words, (ii) setting a minimum cluster frequency of five occurrences, (iii) assigning three labels per cluster, and (iv) employing "walktrap" as the clustering algorithm. Each cluster is represented by a distinct color, and the nodes within each cluster are labeled with keywords associated with the cluster.

**Figure 9.** *Clustering by coupling among the authors.*




Cluster 1: This cluster is characterized by keywords such as "education," "performance," and "university." These keywords suggest a focus on educational performance within academic institutions, with a significant impact indicated by a high global citation score.

Cluster 2: Keywords in this cluster include "students," "instruction," and "knowledge," indicating a focus on student learning and instructional practices. The high centrality and impact scores suggest that research within this cluster has considerable influence in the field.

Cluster 3: This cluster comprises keywords such as "English," "feedback," and "performance," suggesting a focus on language learning and feedback mechanisms. The high impact score indicates that research within this cluster significantly contributes to advancements in these areas.

Cluster 4: Keywords in this cluster include "science," "model," and "students," indicating a focus on scientific education and modeling approaches. The high centrality and impact scores

suggest that research within this cluster has a substantial influence on educational practices related to science.

Cluster 5: This cluster includes keywords such as "feedback," "quality," and "higher education," suggesting a focus on the quality of feedback mechanisms within higher education settings. The absence of an impact score suggests that research within this cluster may be relatively less cited compared to others.

Cluster 6: Keywords in this cluster include "coefficient" and "formative assessment," suggesting a focus on quantitative assessment methods. The moderate impact score indicates that research within this cluster contributes to advancements in assessment practices.

Cluster 7: This cluster comprises the keyword "plagiarism," indicating a focus on academic integrity and plagiarism detection methods. The absence of an impact score suggests that research within this cluster may be less cited compared to others.

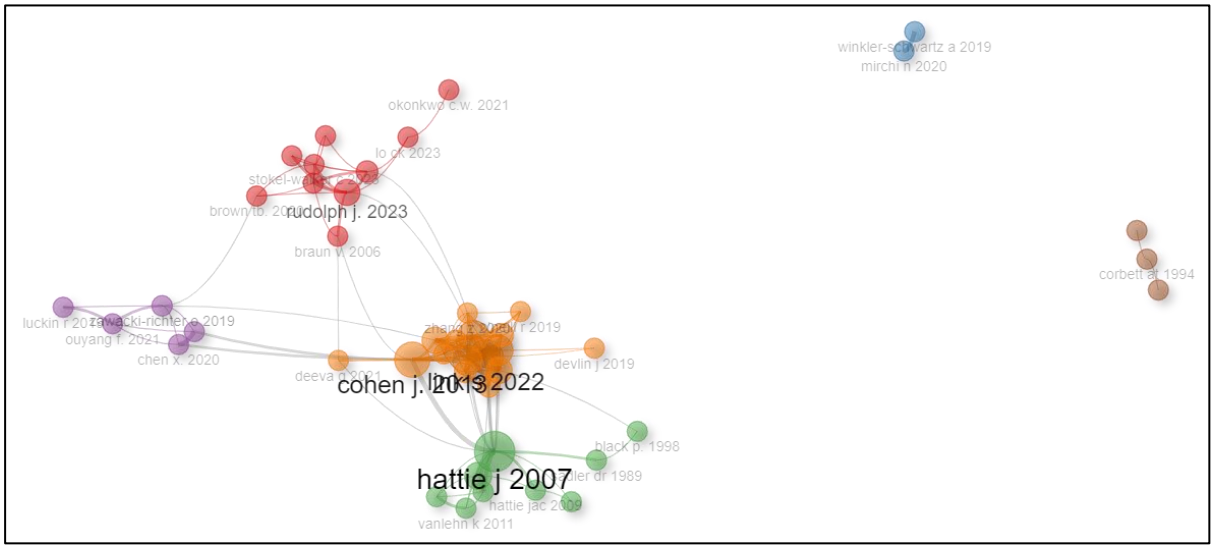
Cluster 8: Keywords in this cluster include "formative assessment," "quality," and "teacher," suggesting a focus on assessment practices and teacher training. The moderate impact score indicates that research within this cluster contributes to advancements in educational assessment.

Cluster 9: This cluster includes the keyword "perceptions," suggesting a focus on understanding learners' perceptions in educational contexts. The absence of an impact score suggests that research within this cluster may be less cited compared to others.

Cluster 10: Keywords in this cluster include "ai" and "curriculum," indicating a focus on integrating artificial intelligence into curriculum development. The moderate impact score suggests that research within this cluster contributes to advancements in AI-based educational technologies.

Furthermore, Figure 10 identifies six clusters with notable works in the field of AI-based feedback tools in education. It reveals distinct clusters of authors based on shared citation patterns, each characterized by unique centrality metrics. The default parameters of the "bibliometrix" package on the web interface "biblioshiny" were employed, including the utilization of the "Walktrap" clustering algorithm.

**Figure 10.** *Co-citation network analysis based on authors.*



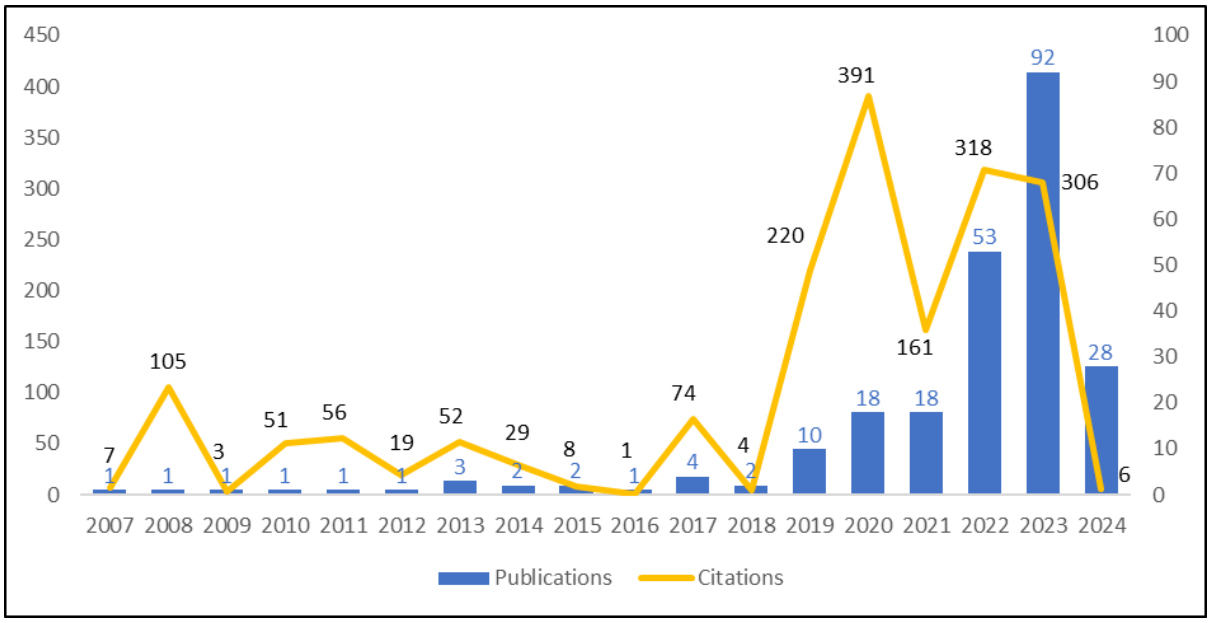

Cluster 1, dominated by recent authors like Rudolph J. and Kasneci E., shows moderate to high betweenness centrality, suggesting their pivotal roles as bridges between other authors. Cluster 2 includes authors such as Mirchi N. and Winkler-Schwartz A., notable for their high closeness centrality, indicating close connectivity within their cluster. Cluster 3, featuring Hattie J. and

Shute VJ, stands out with significant PageRank scores, indicating their substantial citation impact. Clusters 4, 5, and 6 display diverse profiles with authors like Zawacki-Richter O. and Cohen J. demonstrating varying degrees of influence across their respective networks. Overall, these clusters provide insights into the structure and dynamics of scholarly communication within the field, highlighting key authors and their roles in knowledge dissemination and integration.

### 4.4. Comparative Analysis

The chart below provides a comprehensive overview of the evolution of research output and its corresponding impact in the realm of AI-based feedback tools in education, spanning from 2007 to 2024 (see Figure 11). This analysis is particularly insightful when contextualized alongside significant developments in the field, such as the release of ChatGPT, an advanced chatbot developed by OpenAI, launched on November 30, 2022.

**Figure 11.** *Comparison of the number of publications and citations.*




Between 2007 and 2018, the number of publications remained relatively modest, with periodic peaks indicating a gradual but steady accumulation of scholarly work in the domain. However, citation rates during this period varied, with notable spikes observed in 2008, 2011, and 2013. These peaks suggest that despite the limited number of publications, certain research findings garnered substantial attention and recognition within the academic community.

The landscape shifted noticeably post-2018, marked by a significant surge in both the number of publications and their corresponding citations. This trend aligns with the growing interest and investment in AI technologies, including chatbots, for educational purposes. It's worth noting that the release of ChatGPT in late 2022 might have acted as a catalyst for this surge in research activity (Su et al., 2023), contributing to the exponential growth observed in publications and citations in 2022 and 2023.

Between 2019 and 2024, there was an unprecedented surge in research output, marking a period of intense scholarly engagement and innovation within the field. The publication count escalated from 10 in 2019 to a peak of 92 in 2023, showcasing a remarkable expansion of research endeavors. It's worth noting that this study included publications up to February 22, 2024. Given the substantial number of articles published in this short timeframe, it's plausible that the total publication counts for 2024 may surpass that of 2023 by year-end. This surge can be attributed to various factors, including advancements in AI technologies, enhanced

accessibility to research resources, and the growing recognition of the potential of AI-based feedback tools to improve learning outcomes (Sallam et al., 2023).

Simultaneously, the citation rates mirrored this growth trajectory, demonstrating a proportional increase in the impact of research findings during the same period. The surge in citations signifies the growing influence of research in shaping scholarly discourse and informing educational practices, driven by the proliferation of innovative AI-based feedback tools like ChatGPT.

In summary, the comparative analysis underscores the dynamic interplay between research output and impact over time in the field of AI-based feedback tools in education. The release of ChatGPT and other advancements in AI technologies have undoubtedly catalyzed a surge in research activity, shaping the trajectory of scholarly inquiry and innovation for enhancing educational practices.

## 5. CONCLUSION and SUGGESTIONS

The comprehensive analysis conducted on AI-based feedback tools in education provides invaluable insights into the dynamic and evolving landscape of scholarly inquiry within this domain. Through a meticulous examination of publication trends, citation rates, thematic trends, and collaborative dynamics, this study offers a nuanced understanding of the multifaceted nature of research endeavors and the transformative potential of AI technologies in educational settings.

The descriptive analysis serves as a foundational pillar, offering a panoramic view of the scholarly discourse surrounding AI-driven feedback mechanisms. By investigating document characteristics such as publication trends, citation rates, and keyword distributions, this analysis reveals the vitality and relevance of the literature examined. Notably, the exploration of influential countries, affiliations, journals, and publications underscores the global nature of research efforts and the pivotal role of diverse stakeholders in shaping the discourse and advancing knowledge in this field.

Furthermore, the keyword and conceptual analyses provide deeper insights into the prevailing themes and topics within the literature, illuminating the central focus on student performance, feedback provision, and AI technologies in educational contexts. Through co-occurrence networks, the interconnectedness of key concepts and the intricate relationships between different aspects of educational practice and technology utilization are revealed, highlighting the holistic and interdisciplinary nature of research endeavors.

Moreover, the comparative analysis offers a temporal perspective, charting the evolution of research output and impact over time. The exponential growth observed in publications and citations, particularly following significant developments such as the release of ChatGPT, underscores the transformative potential of AI-based feedback tools and the need for continued exploration and innovation in this rapidly evolving field.

Moving forward, it is crucial to address specific gaps in the literature and explore uncharted territories. Future research should focus on investigating the ethical implications of AI-based feedback tools in education, particularly concerning privacy, data security, and potential biases in AI algorithms. Understanding these concerns is essential for developing responsible and equitable deployment strategies.

While AI-based feedback tools offer significant advantages in enhancing educational practices, they also present ethical and social challenges that must be addressed. Concerns around data privacy and security are paramount, as these tools often require the collection and analysis of sensitive student data. Ensuring robust data protection measures and adhering to privacy regulations is crucial to maintain trust and safeguard student information. Additionally, the potential for bias in AI algorithms poses a risk of perpetuating existing inequalities in education. It is essential to critically assess and mitigate biases in AI-driven feedback to ensure fair and

equitable learning opportunities for all students. Addressing these ethical considerations is vital for the responsible deployment of AI technologies in education.

The current study provides a comprehensive bibliometric analysis of AI-based feedback tools in education, revealing significant trends, influential works, and key contributors in the field. While the reliance on the Web of Science database is a limitation, the insights gained are invaluable for understanding the scholarly landscape. Besides, longitudinal studies are imperative to assess the sustained impacts of AI-based feedback tools on student learning outcomes and educational practices over time. By conducting longitudinal research, researchers can better understand how these technologies influence learning trajectories, educational equity, and overall academic achievement.

Additionally, future studies could explore emerging technologies and innovative pedagogical integration strategies to enhance the effectiveness and inclusivity of AI-driven feedback mechanisms. Collaborative efforts across disciplines will be essential in harnessing the full potential of AI technologies to foster positive educational outcomes and address evolving challenges in the field.

## 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.

## Orcid

Mehmet Dönmez https://orcid.org/0000-0003-0339-5135

## REFERENCES

Afzaal, M., Zia, A., Nouri, J., & Fors, U. (2024). Informative feedback and explainable ai-based recommendations to support students' self-regulation. *Technology, Knowledge and Learning, 29*(1), 331–354. https://doi.org/10.1007/s10758-023-09650-0

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Bañeres, D., Rodríguez, M.E., Guerrero-Roldán, A.E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences (Switzerland), 10*(13). https://doi.org/10.3390/app10134427

Barrett, A., & Pack, A. (2023). Not quite eye to A.I.: student and teacher perspectives on the use of generative artificial intelligence in the writing process. *International Journal of Educational Technology in Higher Education, 20*(1), 59. https://doi.org/10.1186/s41239-023-00427-0

Bin-Hady, W.R.A., Al-Kadi, A., Hazaea, A., & Ali, J.K.M. (2023). Exploring the dimensions of ChatGPT in English language learning: a global perspective. *Library Hi Tech*. https://doi.org/10.1108/LHT-05-2023-0200

Bui, N.M., & Barrot, J.S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies,* 1–18. https://doi.org/10.1007/S10639-024-12891-W/TABLES/5

Chang, D.H., Lin, M.P.-C., Hajian, S., & Wang, Q.Q. (2023). Educational design principles of using AI Chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability, 15*(17), 12921. https://doi.org/10.3390/su151712921

Chavez, H., Chavez-Arias, B., Contreras-Rosas, S., Alvarez-Rodríguez, J.M., & Raymundo, C. (2023). Artificial neural network model to predict student performance using nonpersonal

information. *Frontiers in Education, 8*. https://doi.org/10.3389/FEDUC.2023.1106679

Chen, X., Zou, D., Xie, H., Chen, G., Lin, J., & Cheng, G. (2023). Exploring contributors, collaborations, and research topics in educational technology: A joint analysis of mainstream conferences. *Education and Information Technologies, 28*(2), 1323–1358. https://doi.org/10.1007/s10639-022-11209-y

Chen, Z. (2023). Artificial intelligence-virtual trainer: innovative didactics aimed at personalized training needs. *Journal of the Knowledge Economy, 14*(2), 2007–2025. https://doi.org/10.1007/s13132-022-00985-0

Chin, D.B., Dohmen, I.M., Cheng, B.H., Oppezzo, M.A., Chase, C.C., & Schwartz, D.L. (2010). Preparing students for future learning with Teachable Agents. *Educational Technology Research and Development, 58*(6), 649–669. https://doi.org/10.1007/s11423-010-9154-5

Chiu, M.-C., Hwang, G.-J., Hsia, L.-H., & Shyu, F.-M. (2022). Artificial intelligence-supported art education: a deep learning-based system for promoting university students' artwork appreciation and painting outcomes. *Interactive Learning Environments*, 1–19. https://doi.org/10.1080/10494820.2022.2100426

Conrad, E.J., & Hall, K.C. (2024). Leveraging generative AI to elevate curriculum design and pedagogy in public health and health promotion. *Pedagogy in Health Promotion*. https://doi.org/10.1177/23733799241232641

Cowling, M., Crawford, J., Allen, K.-A., & Wehmeyer, M. (2023). Using leadership to leverage ChatGPT and artificial intelligence for undergraduate and postgraduate research supervision. *Australasian Journal of Educational Technology, 39*(4), 89-103. https://doi.org/10.14742/ajet.8598

Cukurova, M., Kent, C., & Luckin, R. (2019). Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *British Journal of Educational Technology, 50*(6), 3032–3046. https://doi.org/10.1111/bjet.12829

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-12402-3

Ebenbeck, N., & Gebhardt, M. (2024). Differential Performance of Computerized Adaptive Testing in Students With and Without Disabilities - A Simulation Study. *Journal of Special Education Technology*. https://doi.org/10.1177/01626434241232117

Elmaoğlu, E., Coşkun, A.B., & Yüzer Alsaç, S. (2024). Digital Transformation: The Role, Potential, and Limitations of ChatGPT in Child Health Education. *American Journal of Health Education, 55*(1), 69–72. https://doi.org/10.1080/19325037.2023.2277937

Farshad, S., Zorin, E., Amangeldiuly, N., & Fortin, C. (2023). Engagement assessment in project-based education: a machine learning approach in team chat analysis. *Education and Information Technologies.* https://doi.org/10.1007/s10639-023-12381-5

Fu, S., Gu, H., & Yang, B. (2020). The affordances of AI-enabled automatic scoring applications on learners' continuous learning intention: An empirical study in China. *British Journal of Educational Technology, 51*(5), 1674–1692. https://doi.org/10.1111/bjet.12995

Gao, R., Merzdorf, H.E., Anwar, S., Hipwell, M.C., & Srinivasa, A.R. (2024). Automatic assessment of text-based responses in post-secondary education: A systematic review. *Computers and Education: Artificial Intelligence, 6,* 100206. https://doi.org/10.1016/j.caeai.2024.100206

Gligorea, I., Cioca, M., Oancea, R., Gorski, A.T., Gorski, H., & Tudorache, P. (2023). Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review. *Education Sciences, 13*(12), 1216. https://doi.org/10.3390/EDUCSCI13121216

Gu, Y. (2024). Research on Speech Communication Enhancement of English Web-based Learning Platform based on Human-computer Intelligent Interaction. Scalable *Computing: Practice and Experience, 25*(2), 709-720. https://doi.org/10.12694/scpe.v25i2.2544

Heeg, D.M., & Avraamidou, L. (2023). The use of Artificial intelligence in school science: a systematic literature review. *Educational Media International, 60*(2), 125–150. https://doi.org/10.1080/09523987.2023.2264990

Hopfenbeck, T.N., Zhang, Z., Sun, S.Z., Robertson, P., & McGrane, J.A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: a perspective article. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1270700

Hopgood, A.A., & Hirst, A.J. (2007). Keeping a Distance-Education Course Current Through eLearning and Contextual Assessment. *IEEE Transactions on Education, 50*(1), 85–96. https://doi.org/10.1109/TE.2006.888905

Jaleniauskienė, E., Lisaitė, D., & Daniusevičiūtė-Brazaitė, L. (2023). Artificial Intelligence in Language Education: A Bibliometric Analysis. *Sustainable Multilingualism, 23*(1), 159–194. https://doi.org/10.2478/sm-2023-0017

Kaldaras, L., Yoshida, N.R., & Haudek, K.C. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education, 7*, 1-15. https://doi.org/10.3389/feduc.2022.983055

Kartal, G., & Yeşilyurt, Y.E. (2024). A bibliometric analysis of artificial intelligence in L2 teaching and applied linguistics between 1995 and 2022. *ReCALL*, 1-17. https://doi.org/10.1017/S0958344024000077

Khoo, E., & Kang, S. (2022). Proactive learner empowerment: towards a transformative academic integrity approach for English language learners. *International Journal for Educational Integrity, 18*(1), 24. https://doi.org/10.1007/s40979-022-00111-2

Kim, M., & Adlof, L. (2024). Adapting to the Future: ChatGPT as a Means for Supporting Constructivist Learning Environments. *TechTrends, 68*(1), 37-46. https://doi.org/10.1007/s11528-023-00899-x

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, … Rummel, N. (2022). Toward learning progression analytics - Developing learning environments for the automated analysis of learning using evidence centered design. *Frontiers in Education, 7*. https://doi.org/10.3389/feduc.2022.981910

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: deep learning really has pedagogical value. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.572367

Lee, A.V.Y. (2023). Supporting students' generation of feedback in large-scale online course with artificial intelligence-enabled evaluation. *Studies in Educational Evaluation, 77*, 101250. https://doi.org/10.1016/j.stueduc.2023.101250

Lee, A.V.Y., Luco, A.C., & Tan, S.C. (2023). A human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educational Technology & Society, 26*(1), 147–159. https://doi.org/10.30191/ETS.202301_26(1).0011

Lee, H.-Y., Chen, P.-H., Wang, W.-S., Huang, Y.-M., & Wu, T.-T. (2024). Empowering ChatGPT with guidance mechanism in blended learning: effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education, 21*(1), 16. https://doi.org/10.1186/s41239-024-00447-4

Li, L., & Kim, M. (2024). It is like a friend to me: Critical usage of automated feedback systems by self-regulating English learners in higher education. *Australasian Journal of Educational Technology, 40*(1), 1–18. https://doi.org/10.14742/AJET.8821

Li, T., Ji, Y., & Zhan, Z. (2024). Expert or machine? Comparing the effect of pairing student teacher with in-service teacher and ChatGPT on their critical thinking, learning performance, and cognitive load in an integrated-STEM course. *Asia Pacific Journal of Education, 44*(1), 45–60. https://doi.org/10.1080/02188791.2024.2305163

Li, W., & Mohamad, M. (2023). An efficient probabilistic deep learning model for the oral proficiency assessment of student speech recognition and classification. International *Journal on Recent and Innovation Trends in Computing and Communication, 11*(6), 411–424. https://doi.org/10.17762/ijritcc.v11i6.7734

Liang, H., Hwang, G., Hsu, T., & Yeh, J. (2024). Effect of an AI-based chatbot on students' learning performance in alternate reality game-based museum learning. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13448

McLaren, B.M., DeLeeuw, KE., & Mayer, R.E. (2011). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education, 56*(3), 574–584. https://doi.org/10.1016/j.compedu.2010.09.019

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R.F. (2020). The virtual operative assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE, 15*(2), e0229596. https://doi.org/10.1371/journal.pone.0229596

Nazari, N., Shabbir, M.S., & Setiawan, R. (2021). Application of artificial intelligence powered digital writing assistant in higher education: randomized controlled trial. *Heliyon, 7*(5), e07014. https://doi.org/10.1016/j.heliyon.2021.e07014

Nimy, E., Mosia, M., & Chibaya, C. (2023). Identifying at-risk students for early intervention-a probabilistic machine learning approach. *Applied Sciences, 13*(6), 3869. https://doi.org/10.3390/APP13063869

Ouyang, F., Wu, M., Zheng, L., Zhang, L., & Jiao, P. (2023). Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. *International Journal of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-022-00372-4

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery, 88*, 1-11. https://doi.org/10.1016/j.ijsu.2021.105906

Palocsay, S.W., & Stevens, S.P. (2008). A study of the effectiveness of web‐based homework in teaching undergraduate business statistics. *Decision Sciences Journal of Innovative Education, 6*(2), 213–232. https://doi.org/10.1111/j.1540-4609.2008.00167.x

Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1255594

Rad, H.S., Alipour, R., & Jafarpour, A. (2023). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: a case of Wordtune application. *Interactive Learning Environments,* 1-21. https://doi.org/10.1080/10494820.2023.2208170

Rahman, M.M., & Watanobe, Y. (2023). ChatGPT for education and research: opportunities, threats, and strategies. *Applied Sciences, 13*(9), 5783. https://doi.org/10.3390/app13095783

Roldán-Álvarez, D., & Mesa, F.J. (2024). Intelligent deep-learning tutoring system to assist instructors in programming courses. *IEEE Transactions on Education, 67*(1), 153–161. https://doi.org/10.1109/TE.2023.3331055

Rosé, C.P., McLaughlin, E.A., Liu, R., & Koedinger, K.R. (2019). Explanatory learner models:

Why machine learning (alone) is not the answer. *British Journal of Educational Technology, 50*(6), 2943–2958. https://doi.org/10.1111/bjet.12858

Rubio-Manzano, C., Lermanda Senoceaín, T., Martinez-Araneda, C., Vidal-Castro, C., & Segura-Navarrete, A. (2019). Fuzzy linguistic descriptions for execution trace comprehension and their application in an introductory course in artificial intelligence. *Journal of Intelligent & Fuzzy Systems, 37*(6), 8397–8415. https://doi.org/10.3233/JIFS-190935

Sağin, F.G., Özkaya, A.B., Tengiz, F., Geyik, Ö.G., & Geyik, C. (2023). Current evaluation and recommendations for the use of artificial intelligence tools in education. *Turkish Journal of Biochemistry, 48*(6), 620–625. https://doi.org/10.1515/tjb-2023-0254

Sallam, M., Salim, N., Barakat, M., & Al-Tammemi, A. (2023). ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J, 3*(1). https://doi.org/10.52225/narra.v3i1.103

Shahriar, S., Allana, S., Hazratifard, S.M., & Dara, R. (2023). A survey of privacy risks and mitigation strategies in the artificial intelligence life cycle. *IEEE Access, 11*, 61829–61854. https://doi.org/10.1109/ACCESS.2023.3287195

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *British Journal of Educational Technology, 50*(6), 3004–3031. https://doi.org/10.1111/bjet.12854

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL,* 1–23. https://doi.org/10.1017/S0958344023000265

Song, P., & Wang, X. (2020). A bibliometric analysis of worldwide educational artificial intelligence research development in recent twenty years. *Asia Pacific Education Review, 21*(3), 473–486. https://doi.org/10.1007/s12564-020-09640-2

Soofi, A.A., & Ahmed, M.U. (2019). A systematic review of domains, techniques, delivery modes and validation methods for intelligent tutoring systems. *International Journal of Advanced Computer Science and Applications, 10*(3), 99–107.

Stojanov, A. (2023). Learning with ChatGPT 3.5 as a more knowledgeable other: an autoethnographic study. *International Journal of Educational Technology in Higher Education, 20*(1), 35. https://doi.org/10.1186/s41239-023-00404-7

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: a framework for applying generative AI in education. *ECNU review of education, 6*(3), 355-366. https://doi.org/10.1177/20965311231168423

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*. https://doi.org/10.1016/j.asw.2023.100752

Téllez, N.R., Villela, P.R., & Bautista, R.B. (2024). Evaluating ChatGPT-generated linear algebra formative assessments. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(5), 75–82. https://doi.org/10.9781/ijimai.2024.02.004

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, 101440. https://doi.org/10.1016/j.tsc.2023.101440

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. pedagogies: *An International Journal, 3*(1), 22-36. https://doi.org/10.1080/15544800701771580

Williams, R.T. (2024). The ethical implications of using generative chatbots in higher education. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1331607

Wong, R.S.Y., Ming, L.C., & Ali, R.A.R. (2023). The intersection of ChatGPT, clinical medicine, and medical education. *JMIR Medical Education, 9*(1), 1-8. https://doi.org/10.2196/47274

Wu, J.-Y., & Tsai, C.-C. (2022). Harnessing the power of promising technologies to transform science education: prospects and challenges to promote adaptive epistemic beliefs in science learning. *International Journal of Science Education, 44*(2), 346–353. https://doi.org/10.1080/09500693.2022.2028927

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology, 00*, 1–17. https://doi.org/10.1111/BJET.13494

Zhang, W., Cai, M., Lee, H. J., Evans, R., Zhu, C., & Ming, C. (2024). AI in medical education: Global situation, effects and challenges. *Education and Information Technologies, 29*(4), 4611–4633. https://doi.org/10.1007/s10639-023-12009-8

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P.L.H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies, 28*(6), 7031–7063. https://doi.org/10.1007/s10639-022-11473-y

Zheng, L., Zhong, L., Niu, J., Long, M., & Zhao, J. (2021). Effects of personalized intervention on collaborative knowledge building, group performance, socially shared metacognitive regulation, and cognitive load in computer-supported collaborative learning. *Educational Technology & Society, 24*(3), 174–193.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 647–659
https://doi.org/10.21449/ijate.1440316
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# A practical guide to item bank calibration with multiple matrix sampling

**Eren Can Aybek**[1*], **Serkan Arıkan**[2], **Güneş Ertaş**[2]

[1]Pamukkale University, Faculty of Education, Department of Educational Sciences, Türkiye
[2]Bogazici University, Faculty of Education, Department of Mathematics and Science Education, Türkiye

**ARTICLE HISTORY**
*Received: Feb. 20, 2024*
*Accepted: Aug. 12, 2024*

**Keywords:**
Multiple matrix sampling,
Item bank development,
Item response theory.

**Abstract:** When it is required to estimate item parameters of a large item bank, Multiple Matrix Sampling (MMS) design provides an efficient way while minimizing the test burden on students. The current study exemplifies how to calibrate a large item pool using MMS design for various purposes, such as developing a CAT administration. The purpose of the current study is to explain and provide an example of how to use MMS design for item bank calibration. Two functions of **mirt** package, `mirt()` and `multipleGroup()` were compared using real data. The results of the present study showed that the standard `mirt()` function is more practical and makes more precise estimations compared to the `multipleGroup()` function.

## 1. INTRODUCTION

Multiple matrix sampling, also known as rotated booklet design or matrix sampling, is a technique where different participants answer different item blocks to reduce the number of items that each examinee answers while ensuring content coverage. This design is based on the idea of dividing a large item pool into blocks of items and administering different but linked booklets to examinees. Therefore, the so-called "item sampling" makes it possible to administer a large set of items (Lord, 1962). The rotation of the items or blocks across the booklets allows us to obtain a reliable and valid measurement of the examinees' abilities as a group and accurate item parameters while reducing the burden of excessive testing. This design is commonly used in international large-scale assessments (ILSAs). The utilization of rotated booklet designs has become increasingly popular in ILSAs, serving as an effective means of gathering population achievement level estimations from a large number of individuals through the use of large item pools. Overall, Multiple Matrix Sampling (MMS) (Lord, 1962; Shoemaker, 1973) allows for calibrating large item pools while minimizing the test burden on students.

The item sampling is termed as the rotated booklet design in large-scale assessments (Rutkowski et al., 2010) or multiple matrix sampling (OECD, 2023). This design is used not only in ILSAs, but in any large-scale assessment that intends to calibrate a large item pool, such as when building an item bank in computerized adaptive testing. As stated by Shoemaker (1973), when the item pool is substantial, the MMS design provides a practical advantage for

*CONTACT: Eren Can AYBEK ✉ erencan@aybek.net Pamukkale University, Faculty of Education, Department of Educational Sciences, Türkiye



**e-ISSN: 2148-7456**

estimating the parameters of the items. Also, this design reduces overall testing time and cost for assessment by reducing testing time per examinee and allowing for a more efficient use of resources (Shoemaker, 1973). Overall, Shoemaker (1973) listed the advantages of MMS as follows: MMS reduces the standard error of the estimate and makes it possible to test a large number of items. Also, as participants answer some parts of the items, testing time is reduced.

Thus, when the purpose is to estimate the proficiency distribution of a population, estimate the person parameters, or estimate item parameters using a large item bank, MMS design provides an efficient way to achieve these goals. Integrating IRT and MMS design allows comparable person or item parameters as IRT can estimate these parameters on a common scale. When estimating population parameters, the latent regression IRT model that utilizes item responses and covariates is a widely used model. In this approach, the multiple imputation technique (Rubin, 1987) is used to estimate the plausible values based on the posterior distributions. When the aim is to estimate person parameters, more items per person are needed to increase the measurement precision of individuals, whereas when the aim is to estimate population parameters, increasing the precision for the population is vital (Gonzales & Rutkowski, 2010). When estimating item parameters, various booklet designs are used. These designs are explained in the following section.

### 1.1. Rotated Booklet Design Types

The requirement to give subtests of items to examinees has prompted the development of various booklet designs. The decision for the specific design is given based on the purpose of the test and the applicability of the design. For computer-based linear tests or paper-based tests, the design needs to be established before finalizing the test booklets. In computerized adaptive tests or multi-stage tests, the items or blocks of items to be administered to examinees are decided based on some algorithms (Gonzales & Rutkowski, 2010).

Gonzales and Rutkowski (2010) categorized booklet designs into complete and incomplete designs. Complete booklet designs are those in which all items or blocks are presented in each form, resulting in all items being answered by all examinees, either in the same order or the rotated order. In complete design, multiple forms can be used by rotating the positions of the items to control the position effect. On the other hand, incomplete booklet designs include booklets that contain a subset of items or blocks. Thus, each examinee answers a subset of all items in the latter one.

Booklet designs are also categorized as balanced and unbalanced designs (Gonzales & Rutkowski, 2010). In a balanced design, every item or block is rotated to appear an equal number of times in each form, whereas in an unbalanced design, some items or blocks rotate, but others generally appear only one time. Balanced booklet designs could control the order effect by counterbalancing.

The balanced incomplete block design (BIBD) was proposed by Lord (1965), in which each subset of items or blocks rotates to appear an equal number of times; therefore, the BIBD balances the position of each item. Table 1 shows one example of a BIBD in which there are a total of 10 items/blocks in the item bank, each student answers five items/blocks, and each item/block appears an equal number of times. On the condition of a large number of items, Shoemaker (1973) investigated the effectiveness of a Partially Balanced Incomplete Block design (PBIBD) compared to a BIBD, finding that the PBIBD could accurately reproduce known means across various conditions. In the PBIBD, each cluster appears a set number of times but does not appear with every other cluster (Rutkowski et al., 2013). A variation of the PIBD was used in TIMSS 2011 and PIRLS 2011.

**Table 1.** *An example of a balanced incomplete block design.*

| Booklet | item1/ block1 | item2/ block2 | item3/ block3 | item4/ block4 | item5/ block5 | item6/ block6 | item7/ block7 | item8/ block8 | item9/ block9 | item10/ block10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | X | X | X | X | X | | | | | |
| 2 | | X | X | X | X | X | | | | |
| 3 | | | X | X | X | X | X | | | |
| 4 | | | | X | X | X | X | X | | |
| 5 | | | | | X | X | X | X | X | |
| 6 | | | | | | X | X | X | X | X |
| 7 | X | | | | | | X | X | X | X |
| 8 | X | X | | | | | | X | X | X |
| 9 | X | X | X | | | | | | X | X |
| 10 | X | X | X | X | | | | | | X |

Table 2 shows one example of an unbalanced incomplete block design (UIBD) in which there are a total of 10 items/blocks in the item bank; each student answers four items/blocks. Items/blocks appear an unequal number of times. Both designs provide links across booklets to calibrate items on the same scale. One of the widely used examples of the BIBD, the BIB7 or Youden squares design, has seven rotated blocks, as shown in Table 3 (Gonzales & Rutkowski, 2010). All blocks are arranged to show up an equal number of times. NAEP, PISA, and TIMSS use designs originated from the BIB7.

**Table 2.** *An example of an unbalanced incomplete block design.*

| Booklet | item1 | item2 | item3 | item4 | item5 | item6 | item7 | item8 | item9 | item10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | X | X | | | | | | | X | X |
| 2 | | | X | X | | | | | X | X |
| 3 | | | | | X | X | | | X | X |
| 4 | | | | | | | X | X | X | X |

**Table 3.** *BIB7 or Youden squares design.*

| Booklet | Blocks | | |
|---|---|---|---|
| 1 | A | B | C |
| 2 | B | C | D |
| 3 | C | D | E |
| 4 | D | E | F |
| 5 | E | F | G |
| 6 | F | G | A |
| 7 | G | A | B |

One commonly used UIBD includes a common part (anchor) and varying blocks. Table 4 depicts an example of such a UIBD, where there is one common block (A) and rotating blocks (B to G). Another version of a UIBD features rotating common parts and non-common parts that appear only once, as depicted in Table 5. In the example given in Table 5, booklet 1 and booklet 2 are linked to each other with C2; booklet 2 and booklet 3 are linked to each other with C3, and so on. For instance, having an item pool of 90 items, the nonrotating part (such as A) might have 10 items, whereas rotating anchors might have 5+5=10 items (such as C1 and C2). Therefore, 90 items could be calibrated while each student answers 20 items in a one-lesson duration.

**Table 4.** *An example of an unbalanced incomplete block design.*

| Booklet | Blocks | | |
|---|---|---|---|
| 1 | B | C | A |
| 2 | C | D | A |
| 3 | D | E | A |
| 4 | E | F | A |
| 5 | F | G | A |
| 6 | G | B | A |

**Table 5.** *An example of an unbalanced incomplete block design.*

| Booklet | Blocks | | |
|---|---|---|---|
| 1 | A | C1 | C2 |
| 2 | B | C2 | C3 |
| 3 | C | C3 | C4 |
| 4 | D | C4 | C5 |
| 5 | E | C5 | C6 |
| 6 | F | C6 | C1 |

### 1.2. Procedural Issues in MMS

Shoemaker (1973) described some procedural issues regarding the application of MMS. The process of MMS consists of three steps: (a) creating booklets with related items or blocks, (b) administering each booklet to selected examinees, and (c) calibrating item and person parameters. Following these steps raises a number of issues to consider when developing the design. For instance, how many subtests will be created? How many test takers are required per booklet? Which is preferable: making fewer booklets with more items in each, or more booklets with fewer items in each? For a more detailed discussion, please visit the book of Shoemaker.

### 1.3. MMS Designs in Large-Scale Assessments

To minimize student burden and estimate population parameters, large-scale assessment programs (e.g., PISA, NAEP, PIRLS, and TIMSS) use the MMS design as they have wide content coverage. As the purpose of these large-scale assessments is to make inferences based on the population, individual scores are not provided to participants. Focusing on population parameters instead of sample parameters allows one to use the most appropriate MMS design based on specific purposes (Gonzales & Rutkowski, 2010).

In PISA 2021, for questionnaire sections, a within-construct matrix sampling design was used. In this design, questions rotate within constructs instead of between constructs. Thus, a student answers different subsets of questions for each construct. In PISA 2018 field trial design, many testlets were used to eliminate the item order effect, and then, students were randomly assigned to these testlets (OECD, 2020). PISA also links their assessments to the one that preceded it by anchor booklets.

TIMSS 2023 administration used a group adaptive assessment design while maintaining the 14-block TIMSS design (Table 6). The booklets were composed of difficult (D), medium (M), and easy (E) items. Seven of the fourteen booklets were created with difficult or medium blocks, whereas the other seven were created with medium or easy blocks. The booklets are linked via common blocks. 70% of the students in high-achieving countries were randomly assigned to more difficult booklets and rest were assigned to the easy booklets (30%); for middle-level countries, these percentages were 50% and 50%; and for low-achieving countries, 30% of the students were randomly assigned to more difficult booklets, and the rest were assigned to the easy booklets (70%). The idea is to better match assessment difficulty with student ability in each country (Yin & Foy, 2021).

**Table 6.** *TIMSS booklet design.*

| Booklets | | Blocks | | | |
|---|---|---|---|---|---|
| More Difficult Booklets | 1 | SM1 | SD1 | MM1 | MD1 |
| | 2 | MD2 | MD3 | SD2 | SD3 |
| | 3 | SM2 | SD2 | MM2 | MD2 |
| | 4 | MD5 | MD1 | SD5 | SD1 |
| | 5 | SM3 | SD3 | MM3 | MD3 |
| | 6 | MM4 | MD4 | SM4 | SD4 |
| | 7 | SD4 | SD5 | MD4 | MD5 |
| Less Difficult Booklets | 8 | ME1 | MM1 | SE1 | SM1 |
| | 9 | SE1 | SE2 | ME1 | ME2 |
| | 10 | ME2 | MM2 | SE2 | SM2 |
| | 11 | SE3 | SE5 | ME3 | ME5 |
| | 12 | ME3 | MM3 | SE3 | SM3 |
| | 13 | SE4 | SM4 | ME4 | MM4 |
| | 14 | ME5 | ME4 | SE5 | SE4 |

First M: Mathematics; Second M: Medium; S: Science; D: Difficult; E: Easy

### 1.4. Studies Based on MMS Designs

MMS designs are used to estimate the proficiency distribution of a population, person parameters, or item parameters utilizing a large item bank. In international large scale assessments, the main purpose of using MMS designs is to estimate population parameters. NAEP, TIMSS, and PISA use MMS design to control the item exposure rate and to ensure that an adequate number of items are presented to each individual for estimating population-level achievement (Rutkowski, 2014). Also, another benefit of using a rotated booklet design is minimizing student burden.

Several studies were conducted to compare different designs using Large Scale Assessment data (e.g., PISA). With a focus on investigating missing data imputation and plausible value generation methodologies, Kaplan and Su (2016) conducted studies to compare three distinct designs: the two-form design, the three-form design, and the PBIBD (partially balanced incomplete block matrix sampling design), utilizing data from the PISA 2012. For a similar purpose, Adam et al. (2013) developed and compared two-form MMS designs using data from the PISA 2006. They have also exemplified the use of MMS designs for questionnaires in their study.

Some studies consider estimating item parameters and population-level parameters for questionnaires. Munger and Loyd (1988) showed that the MMS procedure could be used for the mail survey questionnaires. They reported that the response rate was higher in item-sampled questionnaires. When there are many items in a questionnaire, and the purpose is to estimate item parameters, multiple matrix sampling could be used to minimize the participant burden. In her dissertation, Yan Zhou (2021) conducted a simulation study to develop and compare MMS designs, utilizing non-overlapping short blocks to divide a lengthy context questionnaire (CQ).

Simulation studies provide valuable information about different designs and methods. Gressard and Loyd (1991) conducted a Monte Carlo simulation study to examine how item sampling through item stratification influences parameter estimation when utilizing multiple matrix sampling with achievement data. Gonzales and Rutkowski (2010) compared various designs based on a simulation study. They focus on the effects of various designs on estimating person ability estimates and item parameters and discuss key issues for developing a booklet design. They point out that test developers should find a balanced model for their data since different results would be obtained for the real data.

### 1.5. Present Study

MMS designs are used when a large set of items is required to measure a construct to minimize burden on participants. Like computerized adaptive testing, large scale assessments require a large and calibrated item bank; therefore, the use of rotated booklet design offers advantages in estimating item parameters and developing the item bank. While MMS designs are useful for covering a broad content, minimizing student burden and testing time, and facilitating the estimation of population parameters, estimating item parameters on a common scale requires advanced item analysis techniques. However, the majority of MMS studies focus on estimating student parameters with various designs. Despite the growing number of studies requiring a calibrated large item pool, there is a dearth of literature offering practical guidance on how to estimate item parameters utilizing MMS designs in real datasets. Thus, the purpose of the current study is to explain and provide an example of how to calibrate a large item bank that is given to students with an MMS design. In the current study, it is exemplified how a real item pool, including 540 math items at the fourth-grade level can be calibrated via UIBD.

## 2. METHOD

### 2.1. Participants

The current study makes use of items and data from a project that aims to develop a CAT system for fourth graders. In the field test phase, 3108 students- 66% of public schools and 34% of private schools-participated in order to calibrate an item bank including 540 mathematics items. A total of twelve public schools and twenty-three private schools participated in the current research. The schools and the students volunteered to attend the study.

### 2.2. Instrument

To create a computerized adaptive test system, first, a large item pool of fourth-grade mathematics items, 540 items, was developed. These items were developed based on TIMSS assessment framework where items were planned to measure three types of cognitive dimensions: knowing, applying and reasoning (Mullis et al., 2021). Due to the hierarchical nature of TIMSS taxonomy, knowing items are supposed to be simpler than applying items, whereas reasoning items are the most cognitively demanding. To enable simultaneous calibration of these 540 items, they were placed into 36 booklets, each containing 20 items (see Table 7). Items were placed accordingly to create parallel booklets in terms of content and cognitive dimensions, and applying items were mainly placed to anchor items as applying items are suitable to the majority of the students. This procedure has been done by measurement specialists and math educators according to the test blueprint. Using blocks by grouping items was also useful to maintain the similarity of the item contexts for each booklet. Otherwise, participants' scores could be affected by unequal context distribution, and this situation might create construct-irrelevant variance (Gonzales & Rutkowski, 2010).

The testing time is one of the most significant limitations in actual data collection. Considering that classes often run 40 or 50 minutes, 20 items per student would be considered sufficient. Thus, a UIBD was selected in order to calibrate 540 items while administering the minimum item per student. Complete booklet designs were not selected as they required 540 items to be given to each pupil. Furthermore, the BIBD were not preferred since they necessitated using an equal quantity of each item, which meant making more booklets. For instance, a BIBD with 20 items per a booklet will result in 540 booklets; a very large sample size is needed to calibrate that many booklets. Therefore, to have a minimum number of booklets, a UIBD was selected. In the UIBD, similar to the one in Table 5, 540 items could be calibrated using 36 booklets. In the current study design, the first blocks, like block As, had 10 items, and the anchor blocks, block Bs and block Cs, each had five items. Therefore, we end up with a total of 20 items per booklet and 36 booklets. Booklet 1 is linked to booklet 2 via B1 and to booklet 36 via C18; booklet 2 is linked to booklet 1 via B1 and to booklet 3 via C1, and so on. The total quantity of

booklets will differ based on the number of items in each block; for instance, fewer booklets will be produced overall if there are fewer items in the anchor blocks and more items in the initial blocks. But since fewer items in anchor blocks could raise the standard error, a substantial number of items are needed in anchor blocks.

**Table 7.** *Multiple Matrix Design of the current study.*

| | Unique Items Blocks A (36 Blocks; 10 items each) | Anchor Item Blocks B (18 Blocks; 5 items each) | Anchor Item Blocks C (18 Blocks; 5 items each) |
|---|---|---|---|
| Booklet 1 | Block A1 (items 1-10) | Block B1 (items 361-365) | Block C18 (items 536-540) |
| Booklet 2 | Block A2 (items 11-20) | Block B1 (items 361-365) | Block C1 (items 451-455) |
| Booklet 3 | Block A3 (items 21-30) | Block B2 (items 366-370) | Block C1 (items 451-455) |
| Booklet 4 | Block A4 (items 31-40) | Block B2 (items 366-370) | Block C2 (items 456-460) |
| Booklet 5 | Block A5 (items 41-50) | Block B3 (items 371-375) | Block C2 (items 456-460) |
| Booklet 6 | Block A6 (items 51-60) | Block B3 (items 371-375) | Block C3 (items 461-465) |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| Booklet 33 | Block A33 (items 321-330) | Block B17 (items 441- 445) | Block C16 (items 526-530) |
| Booklet 34 | Block A34 (items 331-340) | Block B17 (items 441- 445) | Block C17 (items 531-535) |
| Booklet 35 | Block A35 (items 341-350) | Block B18 (items 446- 450) | Block C17 (items 531-535) |
| Booklet 36 | Block A36 (items 351-360) | Block B18 (items 446- 450) | Block C18 (items 536-540) |

## 2.3. Data Analysis

Student data was collected on the Concerto Platform as a long data format (examinees in rows and variables in columns). Data cleaning and preparations were handled using R (R Core Team, 2023) and the *dplyr* package (Wickham et.al., 2023). Following the administration of the booklets, four items were removed from the dataset as two items had zero variances, and the other two items had a printing error. Then the local independence assumption was checked by using Yen's *Q3* statistic with a 0.20 cut-off criterion (Chen & Thissen, 1997). According to Yen's *Q3* statistics, 23 items that violate local independence assumption were eliminated.

Items were calibrated with the *mirt* package (Chalmers, 2012) using `mirt()` and `multipleGroup()` functions. We refer to the method that uses `mirt()` function as the standard method and `multipleGroup()` function as the multiple group method. The standard method is used for IRT item calibrations according to dichotomous and polytomous IRT models. On the other hand, the multiple group method is utilized for vertical scaling (particular items answered by only one group while both groups answered common anchor items) in addition to its major applications, such as detecting differential item functioning (DIF) and differential test functioning (DTF). It divides the data into subsets, applies the conventional procedure to each subset independently, and then aggregates the outcomes. During this process, multiple group method allows the user to constrain some parameters to be equal (e.g., anchoring). On the other hand, the standard method uses the entire dataset, assigns plausible values to missing data, and then makes the calibrations (Chalmers, 2023).

For the `multipleGroup()` function, booklets were used as the grouping variable. However, because of the `multipleGroup()` function's massive processing power needs, it is typically necessary to perform the estimations as paired pairs in order to estimate the standard errors of item parameter estimates. That's why we run the `multipleGroup()` function for paired booklets: booklet 1 and booklet 2; booklet 2 and booklet 3; booklet 3 and booklet 4, and so on.

Despite the enormous overall number of students in the current study, there were around 90 pupils per booklet. Therefore, the Rasch model was selected to calibrate the item bank using both methods (O'Neill et.al., 2020). Then, the difficulty (b) parameters and their standard errors

for both methods were compared.

In order to evaluate the consistency of b parameters, the correlation between IRT b parameters and Classical Test Theory (CTT) *p* statistics were estimated. Research showed that under the CTT and IRT frameworks, there is a strong correlation between item difficulty parameters (MacDonald & Paunonen, 2002). The significance of the difference between these correlations obtained from both calibration methods was tested by using Fisher's Z test, and Cohen's *q* statistics for the effect size. The calculations for the Fisher's Z test and Cohen's *q* statistics were handled with the *diffcor* package (Blotner, 2024) in R. The R codes used in the data analysis can be reached through https://github.com/ecaybek/rbd

## 3. FINDINGS

### 3.1. Comparison of b Parameters

The item difficulty parameters were calibrated using the Rasch model, and descriptive statistics for the b parameters are presented in Table 8 for the multiple group method and standard method. The results showed that the item bank covered an ability range of -4.66 to 2.90 for the multiple group method and -4.62 to 2.88 for the standard method. The mean of the b parameters for both methods were close to zero and b parameters were normally distributed according to both methods.

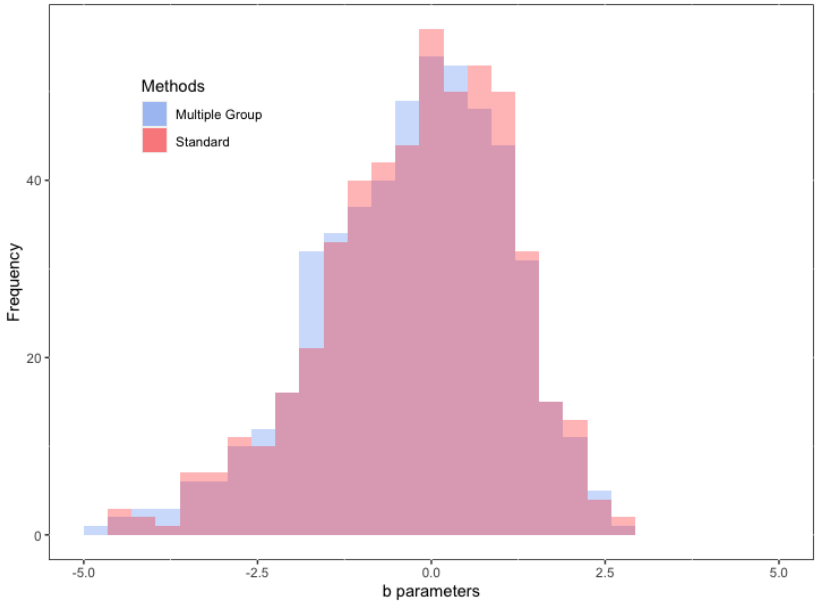
**Table 8.** *Descriptive statistics of b parameter estimations by two methods.*

| Methods | *k* | min | max | mean | median | *s* | skewness | kurtosis |
|---|---|---|---|---|---|---|---|---|
| Multiple Group | 513 | -4.66 | 2.90 | -0.20 | -0.13 | 1.36 | -0.52 | 0.10 |
| Standard | 513 | -4.62 | 2.88 | -0.21 | -0.08 | 1.35 | -0.54 | 0.17 |

*k: number of items; s: standard deviation*

The mean difference of b parameters between the two methods was not significant ($t_{1024}$ = -0.90; $p$ = .37). The distribution of the b parameters for the multiple group and the standard method is presented in Figure 1. As can be seen in Table 8 and Figure 1, according to the estimations from both methods, the item bank had items targeting a very large range of ability levels, especially for very low ability levels (lower than -2) and high ability levels (higher than 2).

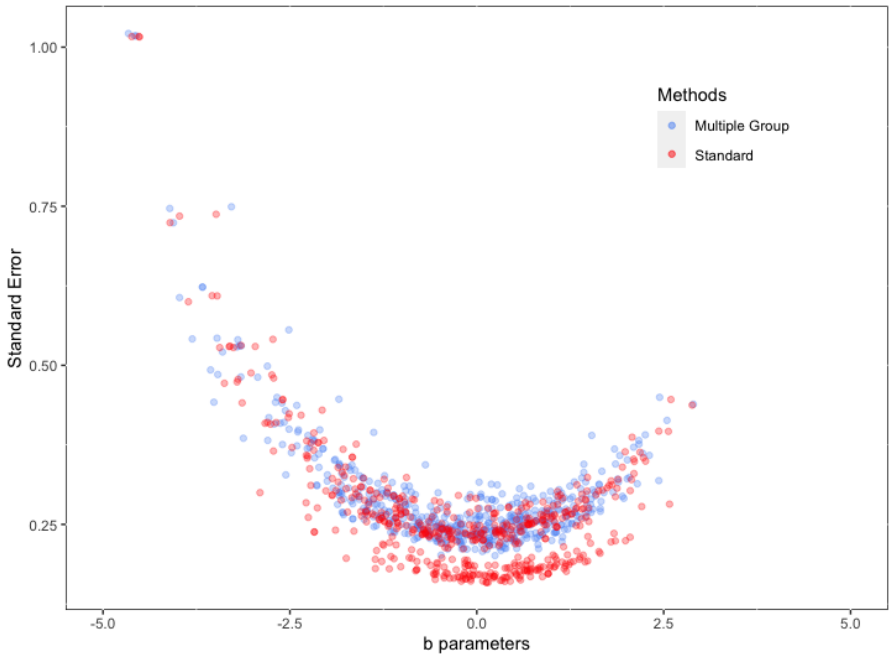
**Figure 1.** *Distribution of the b parameters of the item bank.*



### 3.2. Evaluation of Standard Errors

The standard errors of the b parameters estimated by both methods were compared to gain a better understanding of the item parameter estimations (see Figure 2). The standard method tends to estimate b parameters with smaller standard errors than the multiple group method. This discrepancy may be due to the `multipleGroup()` function's enormous processing power requirements.

**Figure 2.** *Distribution of the SEs of the b parameters of the item bank.*



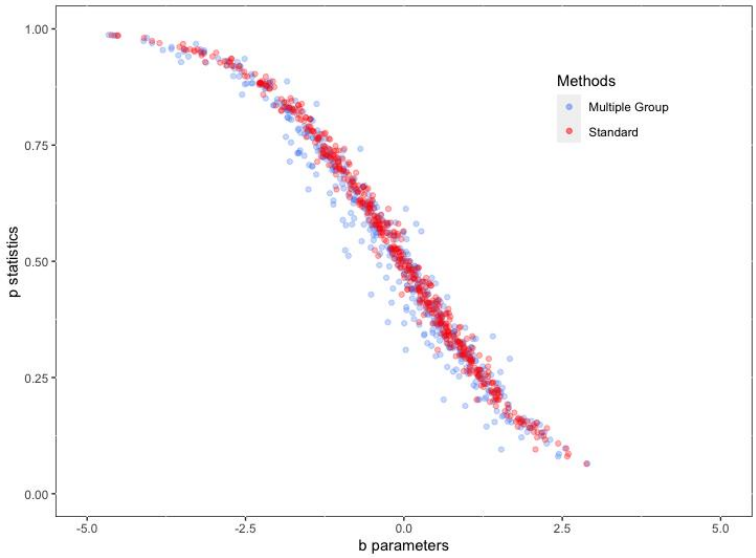

Because of this need, b parameters are estimated by using booklet pairs (Booklet 1 - Booklet 2, Booklet 2 - Booklet 3, and so on) with `multipleGroup()` function. Because the multiple group methodology used data from booklet pairs, while the standard method used the complete dataset, the multiple group method likely estimated the b parameters with higher standard errors due to the smaller dataset size. The U-shaped plot of the standard error occurs due to relatively easy and difficult items having fewer observations for estimating the lower asymptote (Thissen & Wainer, 1982). We believe that the items at the tails have very similar standard errors for both methods.

### 3.3. Correlation among IRT and CTT Difficulty Parameters

It is also important to evaluate the correlation between IRT b and the CTT *p* statistics. Since CTT *p* statistics and the IRT b parameters are related to the area under the normal distribution curve, this investigation provided us insight into how well the two methods estimated the item parameters. Thus, the scatter plot between the IRT b parameters and the CTT *p* statistics for both methods is shown in Figure 3.

The scatter plot shows that the IRT b parameter estimates from both methods highly correlate with the CTT *p* statistics. On the other hand, the standard method has a stronger relationship with the CTT *p* statistics. While the correlation coefficient between multiple groups and the CTT was found to be -0.972, the correlation coefficient between the standard method and the CTT was found to be -0.981. Fisher's Z test showed that the standard method had a significantly higher correlation with the CTT *p* statistics than the multiple group method ($z = 3.059$; $p < .01$). On the other hand, Cohen's *q* was found to be 0.19, which indicates the size of the difference was small (Cohen, 1988).

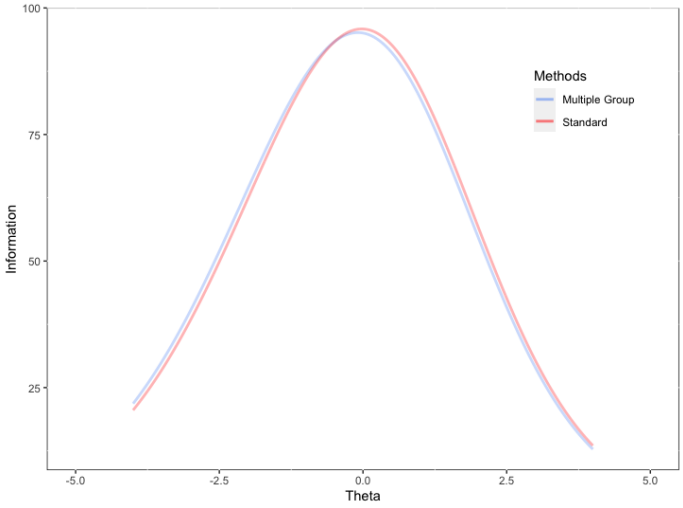
**Figure 3.** *Scatter plot of the b parameters and the p statistics.*




## 3.4. Comparison of Test Information Functions

Finally, the test information functions of the item bank were drawn using both methods (Figure 4), and both methods generated very similar test information functions. To sum up all the findings, the standard method has been found more efficient by the manner of computing power and simplicity while there were no significant differences between mean b parameter estimations; the standard method estimates the b parameters with smaller standard error and higher correlation with the CTT *p* statistic.

**Figure 4.** *Information functions of the item bank.*




## 4. DISCUSSION and CONCLUSION

The current study aims to exemplify how to calibrate an item bank utilizing MMS design for various purposes, such as developing a CAT administration. Therefore, the current study focuses on why, when, and how to use MMS design. Studies on MMS mostly focus on estimating student parameters, and to the best of our knowledge, estimating item parameters in MMS designs is not prevalent in the literature. Thus, there is a need to demonstrate how to calibrate a large item bank using Multiple Matrix Sampling. Calibrating a large item pool requires deciding on a specific booklet design by considering methodological and practical issues. As Gonzalez and Rutkowksi (2010) stated, in any design, there is a trade-off between what is desired and what is practical based on the purpose of the assessment and existing

resources. More items mean more precision; however, it is more laborious. Integrating the benefits of IRT and MMS, it is more practical and efficient to estimate item parameters of large item pools. Given the constraints of data collection, such as class time of schools and low stake consequences of data collection for participants, it is a kind of must to administer a relatively restricted number of items to students. Depending on the topic, student level and cognitive load, 15 to 20 items may be ideal to administer in a single course time.

In the current study, items (4th grade mathematics) were developed based on the TIMSS Assessment Framework. TIMSS fourth-grade mathematics assessment included three content domains: (1) number, (2) measurement and geometry, (3) data, and three cognitive domains: (1) knowing, (2) applying, (3) reasoning. A substantial number of items within each category should have been administered to enable precise estimation of proficiency distribution (Rutkowski et al., 2013). A total of 540 items were developed in this study. Obviously, it was impossible to administer every item to all examinees. One of the appropriate models to calibrate these items was an unbalanced incomplete booklet design. Thus, in a single lesson period, each student encountered 20 items from all content and cognitive areas.

As simulation studies provided somewhat clean results, using real data from a test provides valuable information and is important for sharing the experience. As Gonzales and Rutkowski (2010) stated, test developers should find a balanced model for their data since different results would be obtained for the real data. Thus, the current study explained the procedures and challenges of calibrating a large item pool using real data.

Each item in the current study was responded to by a varying number of participants due to the design and challenges in reaching out to a big sample. With 36 booklets and 540 items to calibrate, anchor items were answered by approximately 180 students, while non-anchor items were answered by approximately 90 students. As a result, the mean standard error of anchor items was smaller than non-anchor items. In a balanced design, the number of students per item for both anchor and non-anchor items would be similar, resulting in similar standard errors. However, balanced designs will result in more booklets, which require more pupils.

The standard errors of item difficulties were higher for items with extreme difficulties. The estimates of difficulty for items that were very easy and very difficult were less precise compared to the items with medium level difficulty. Gonzalez and Rutkowksi (2010) also reported a similar finding and reported that having more people responding to the items, the precision increases, especially for the extremes. On the one hand, this is a predictable outcome; an item bank for a CAT administration necessitates a huge number of extreme items in order to adequately match student abilities.

The *mirt* package provides very useful tools not only for the conventional item bank development process but also for item bank development under the MMS design. The package includes two functions, `mirt()` and `multipleGroup()`, which are very useful for MMS design. The results of the present study showed that the standard `mirt()` function is more practical and makes more precise estimations when it is compared to the `multipleGroup()` function. It is practical because when `multipleGroup()` function was used with booklet pairs, the estimations took around 42 seconds, while `mirt()` function estimated the item parameters in around 24 seconds. Moreover, the `multipleGroup()` function was incapable of calculating standard errors when 36 booklets were simultaneously included in the analysis. The standard error estimation failed with support not only from the personal computers of the researchers but also from Google Cloud servers. Even though there was no significant difference between the mean of b parameter estimations from both methods, `mirt()` function also estimated the b parameters with less standard error and showed higher correlation with the CTT *p* statistics.

Overall, comparing the multiple group method and standard method, while there were no statistically significant differences between the mean b parameter estimations, the standard method was found to be more efficient in terms of computing power and simplicity. It also estimates b parameters with a smaller standard error and a higher correlation with the CTT *p* statistic.

### 4.1. Further Suggestions and Limitations

For practical researchers, the standard `mirt()` function is more useful and precise than the `multipleGroup()` function for calibrating item banks with the MMS design. Also, a simulation study can be conducted to compare the bias and RMSE values of the b parameter estimations from both methods. Counterbalancing could also be used to minimize the effect of item order.

One limitation of the current study is that the Rasch model was used to evaluate item discrimination. Due to sample size per booklet, the Rasch model was chosen. A larger sample size per booklet would be better to test the other IRT models. Another limitation is the pairing of booklets when making calibrations via `multipleGroup()` function due to its computational requirements. It would be good to compare the results of this function by running without pairing the booklets.

### Acknowledgments

This study was supported by Bogazici University Scientific Research Commission (Project no: BAP-SUP 17002). The preliminary results were presented in 1st Adaptive Test Research National Symposium in 14th - 15th September, 2023 in Bogazici University, Istanbul, Türkiye.

### 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**: Bogazici University, 84391427-050.01.04-E.9920.

### Contribution of Authors

**Eren Can Aybek:** Research Design, Methodology, Data Collection, Data Analysis, and Writing. **Serkan Arıkan:** Literature Review, Research Design, Methodology, Data Collection, Supervision, Writing and Critical Review. **Güneş Ertaş:** Literature Review, Methodology, Writing, and Data Collection.

### Orcid

Eren Can Aybek https://orcid.org/0000-0003-3040-2337
Serkan Arıkan https://orcid.org/0000-0001-9610-5496
Güneş Ertaş https://orcid.org/0000-0001-8785-7768

### REFERENCES

Blötner, C. (2024). *Package 'diffcor'*. https://cran.r-project.org/web/packages/diffcor/diffcor.pdf

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433–448). Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Chalmers, R.P. (2012). mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software*, *48*, 1-29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R.P. (2023). *Package "mirt"*. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Gonzalez, E., & Rutkowski, L. (2010). *Principles of Multiple Matrix Booklet Designs and Parameter Recovery in Large-Scale Assessments* (pp. 125–156). IERI.

Gressard, R.P., & Loyd, B.H. (1991). A comparison of item sampling plans in the application of multiple matrix sampling. *Journal of Educational Measurement*, *28*(2), 119–130.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, *41*(1), 57–80.

Lord, F.M. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, *22*(2), 259–267. https://doi.org/10.1177/001316446202200202

Lord, F.M. (1965). Item sampling in test theory and in research design. *ETS Research Bulletin Series*, *1965*(2), i–39. https://doi.org/10.1002/j.2333-8504.1965.tb00968.x

Macdonald, P., & Paunonen, S.V. (2002). A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, *62*(6), 921-943. https://doi.org/10.1177/0013164402238082

Munger, G.F., & Loyd, B.H. (1988). The use of multiple matrix sampling for survey research. *The Journal of Experimental Education*, *56*(4), 187–191.

OECD. (2020). *PISA 2018 Technical Report-PISA*. OECD Publishing, Paris. Retrieved from https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2023). *PISA 2022 Technical Report-PISA*. OECD Publishing, Paris. Retrieved from https://www.oecd.org/pisa/data/pisa2022technicalreport/

O'Neill, T.R., Gregg, J.L., & Peabody, M.R. (2020). Effect of sample size on sommon item equating using the dichotomous rasch model. *Applied Measurement in Education*, *33*(1), 10–23. https://doi.org/10.1080/08957347.2019.1674309

Rubin, D.B. (2009). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education*, *27*(2), 115-132. https://doi.org/10.1080/08957347.2014.880440

Rutkowski, L., Gonzalez, E., Davier, M. von, & Zhou, and Y. (2013). Assessment design for international large-scale assessments. In *Handbook of International Large-Scale Assessment*. Chapman and Hall/CRC.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in Secondary Analysis and Reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Shoemaker, D.M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Ballinger Publishing Company.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4).

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science*. O'Reilly Media, Inc.

Yin, L., & Foy, P. (2023). *TIMSS 2023 Assessment Design*. In I.V.S. Mullis, M.O. Martin, & M. von Davier (Eds.), TIMSS 2023 Assessment Frameworks. Boston College, TIMSS & PIRLS International Study Center.

Zhou, Y. (2021). *Improving Multiple Matrix Sampling Design for Questionnaires*. Indiana University.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 660–674
https://doi.org/10.21449/ijate.1384824
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Investigating the quality of a high-stakes EFL writing assessment procedure in the Turkish higher education context

**Elif Sarı** [1*]

[1]Karadeniz Technical University, School of Foreign Languages, Trabzon, Türkiye

**ARTICLE HISTORY**

*Received: Nov. 01, 2023*

*Accepted: Aug. 26, 2024*

**Keywords:**

EFL writing assessment,
Writing assessment in higher education,
Scoring variability,
Scoring reliability,
Generalizability (G-) theory.

**Abstract:** Employing G-theory and rater interviews, the study investigated how a high-stakes writing assessment procedure (i.e., a single-task, single-rater, and holistic scoring procedure) impacted the variability and reliability of its scores within the Turkish higher education context. Thirty-two essays written on two different writing tasks (i.e., narrative and opinion) by 16 EFL students studying at a Turkish state university were scored by 10 instructor raters both holistically and analytically. After the raters completed the scoring procedure, semi-structured individual interviews were held with them to gain insight into their views regarding the quality of the current scoring procedure. The G-theory results showed that the reliability coefficients obtained from the current scoring procedure would not be sufficient to draw sound conclusions. The quantitative results were partly supported by the qualitative data. Important implications were discussed to improve the quality of the current high-stakes EFL writing assessment policy.

## 1. INTRODUCTION

Reliability and validity are the two fundamental components of assessment. Reliability refers to the consistency of scores obtained across a range of circumstances and conditions (Johnson et al., 2009). Without consistency, it becomes challenging to draw meaningful conclusions or make accurate inferences about an individual's true ability. Validity, as the other important concept in assessment, refers to the degree to which an assessment tool accurately measures what it claims to measure (Bachman, 1990). It means that validity is “the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (Messick, 1989, p. 39). If a given score is not valid, that would impair the fairness of the judgment made about the test takers’ performance (Kane, 2010). Although consistency in test scores does not necessarily ensure validity, it is a fundamental requirement for it (Popham, 1981). Consequently, reliability is viewed "as a cornerstone of sound performance assessment" (Huang, 2008, p. 202).

It is necessary to ensure the reliability and fairness of scores in any assessment procedure, especially when the decisions made on these scores significantly impact students' lives (AERA, APA, & NCME, 2014). However, it is difficult to provide consistency among or within raters due to a variety of rater differences, such as educational background, linguistic background,

*CONTACT: Elif SARI ✉ elifsari@ktu.edu.tr Karadeniz Technical University, School of Foreign Languages, Trabzon, Türkiye



**e-ISSN: 2148-7456**

professional experience, and beliefs and expectations (Huot, 1990). The factors impacting the reliability and fairness of scores in ESL/EFL writing assessment can be categorized under three headings: 1) the factors related to the rater, 2) the factors related to the writing task, and 3) the factors related to the scoring method (Barkaoui, 2007; Barkaoui, 2008; Gebril, 2009; Huang, 2011; Weigle, 2002).

The literature has shown that rater-related factors such as the rater's native language (Cheong, 2012; Kim & Gennaro, 2012; Shi, 2001), professional experience (Barkaoui, 2010; Rinnert & Kobayshi, 2001; Şahan & Razı, 2020), professional background (Elorbany & Huang, 2012; Weigle, Boldt, & Valsecchi, 2003), and training (Attali, 2020; Fahim & Bijani, 2011; Weigle, 1994) affect the scoring variability and reliability. Several studies indicated that native English-speaking (NES) raters exhibited different scoring tendencies from non-native English-speaking (NNES) raters. Shi (2001) discovered that NES raters tended to exhibit a more favorable disposition when scoring content and language aspects, whereas NNES raters showed a tendency to be critical, particularly regarding organization and essay length. Similarly, in Kim and Gennaro's (2012) research, NNES raters were inclined to be more severe and displayed more variability in their scoring compared to NES raters. In contrast, Cheong (2012) observed that NES raters awarded lower grades and applied stricter evaluation criteria across three domains: content, organization, and language use. Regarding the impact of raters' professional experience on their scores, Rinnert and Kobayashi (2001) concluded that the least experienced Japanese raters gave higher scores compared to NES raters, and the groups differed in the criteria they prioritized. Barkaoui (2010) found that when employing holistic and analytic scales, experienced and inexperienced raters exhibited varying degrees of severity and leniency. Novice raters tended to be more lenient in their ratings compared to experienced raters. In addition, Şahan (2018) observed that highly experienced raters were more lenient and assigned higher scores, particularly for low-quality essays. To investigate how raters' professional backgrounds impact their scoring behaviours, Weigle, Boldt, and Valsecchi (2003) studied how instructors from different professional backgrounds evaluate text-responsible writing by ESL students. They found that raters from different disciplines had varying assessments, with English department raters being the strictest and history department raters being the most lenient. The study also revealed that English department raters placed more emphasis on grammar. In a separate study by Elorbany and Huang (2012), it was observed that raters with different educational backgrounds displayed different assessment behavious. Teacher candidates majoring in TESOL provided more consistent scores compared to the raters who did not have a TESOL background. To reveal the impact of rater training on raters' scores, Weigle (1994) studied experienced and inexperienced raters' scoring behaviours before and after they received training and revealed that inexperienced raters' scoring behaviours changed after training while the others gave similar scores both before and after the training. Similarly, Fahim and Bijani's (2011) study found that providing training to raters improved self-consistency and reduced severity and bias in the rating process. Finally, Attali (2020) compared inexperienced and experienced raters and found their ratings to be similar after initial training, but inexperienced raters showed more score variability.

Several studies indicated that writing task (e.g., narrative, persuasive, etc.) is another factor that affects the scoring variability and reliability (Cumming et al., 2002; Gebril, 2009; Hamp-Lyons & Mathias, 1994; Weigle, 1999; Zhao & Huang, 2020). For instance, as Hamp-Lyons and Mathias (1994) discovered, essays written in response to challenging writing prompts were given higher scores than those written in response to easy prompts. They also discovered that the category that the raters considered the simplest received the lowest ratings, whereas the category perceived as the most challenging received the highest ratings. In a similar vein, Weigle (1999) found that inexperienced raters assigned lower grades to certain essay types compared to experienced raters, but training reduced the differences. Cumming et al. (2002) also observed that writing tasks influenced raters' scoring processes and their focus on different

essay features. Additionally, Gebril (2009) and Zhao and Huang (2020) showed that including different task types increased scoring reliability.

The scoring method used by raters also affects the score variability and reliability in writing assessments. Therefore, several studies were undertaken to investigate how holistic and analytic scoring methods impact the variability and reliability of scores (Barkaoui, 2007, 2010; Han, 2013; Liu & Huang, 2020; Song & Caruso, 1996). For instance, in their study, Song and Caruso (1996) compared the holistic and analytic scoring of compositions written by native and non-native English speakers and found no statistically significant difference between the groups stemming from the rating method. Barkaoui (2007) investigated how different scoring methods impacted EFL essays and found higher inter-rater reliability with holistic rating. In a later study, Barkaoui (2010) examined the influence of the rating method on writing evaluation and found that the rating method significantly impacted the raters' scoring processes and the writing aspects they prioritized. In the same vein, Han's (2017) study suggested that detailed training made holistic scoring as reliable as analytic scoring. More recently, Liu and Huang (2020) evaluated the scoring policy of a standardized EFL assessment in China and showed that analytic scoring produced more reliable scores. It also showed that scoring reliability could improve with the increased number of tasks.

To sum up, the research has indicated that ESL/EFL writing assessment is a problematic issue as it is essential to control several factors that impact the variability, reliability, and thus the fairness of scores. In this sense, it is crucial to investigate the variability and reliability issues in any writing assessment procedure that is used to make critical judgments about the examinees' writing abilities (AERA, APA, & NCME, 2014). For example, in Turkish higher education, students' writing performance is assessed to make some high-stakes decisions such as determining students' language proficiency when they are enrolled in the departments that are related to English Language Teaching or Literature or selecting students who will take part in the international exchange programs like Erasmus$^{+}$. Although each university conducts its own writing assessment procedure, students' writing performance is mostly assessed using a single-task, single-rater, and holistic scoring procedure as it is considered to be more time-efficient and cost-effective. Since the studies reviewed above were mostly conducted in different writing assessment contexts, there is limited information regarding the scoring variability and reliability of the writing assessment procedures employed specifically in the context of Turkish higher education. Therefore, it becomes imperative to undertake an in-depth exploration of the quality of writing assessment within this specific educational context. To bridge this existing gap in the literature, this study set out to evaluate the quality of a single-task, single-rater, and holistic scoring method within the Turkish higher education context, focusing on its potential effects on scoring variability and reliability using the G-theory framework. Studying the variability and reliability of this institutional writing evaluation process can have significant implications for assessment policymakers in this specific context (i.e., the Turkish higher education context) as it helps them determine the optimal approach for a high-quality writing assessment procedure, focusing on key factors such as the number of tasks, the number of raters, and the scoring method. Furthermore, the implications are far-reaching and extend to professionals engaged in the evaluation of EFL writing skills on a global scale. Consequently, the findings and insights generated by this study could substantially inform and enhance the practices and policies of assessment professionals and policymakers alike, with the potential to foster improvements not only in Turkish higher education but also in the broader context of EFL writing assessment. The study was directed by four specific research questions, which are as follows:

1. What are the sources of variability in scores given to the EFL papers?
2. How reliable are the EFL scores in terms of G-coefficients for norm-referenced interpretation and dependability coefficients for criterion-referenced score interpretations?

3. Does the scoring reliability change when the number of raters, tasks and the scoring method change?

4. What are the raters' views regarding the overall quality of the single-task, single-rater, and holistic writing assessment procedure?

### 1.1. G-theory Framework

Classical Test Theory (CTT), the conventional measurement model, posits that a measured score (X) comprises a true score (T) and an error score (E). The true score is the test-takers' actual performance resulting from their ability, while the observed score reflects the interaction between the true score and the error score, which are influenced by some external factors apart from the ability intended to be measured (Fulcher & Davidson, 2007). CTT primarily considers two sources of error (i.e., a single ability and a single source of errors), while G-theory recognizes that the sources of error in measurement are diverse and can come from various facets or components. (Bachman, 1990; Briesch et al., 2014). These sources, commonly known as facets, can include different raters, items, occasions, or any other factors that contribute to measurement variability. By incorporating these facets into the analysis, G-theory provides a more detailed understanding of how these different sources impact the reliability and generalizability of the obtained scores. (Shavelson & Webb, 1991).

The G-theory analysis includes two phases: the generalizability study (G-study) and the decision study (D-study). The G-study focuses on assessing the generalizability, or the extent to which the obtained results can be applied beyond the specific conditions of the study. It aims to estimate the various sources of error in measurement and to determine how they contribute to the variability of scores. By examining different facets of measurement, such as raters, tasks, and occasions, the G-study helps researchers understand the factors that affect the reliability and validity of the measurement instrument or procedure (Barkaoui, 2007; Huang et al., 2014). The D-study, on the other hand, is a phase that focuses on making decisions using the measurement data revealed in the G-study. By utilizing the results from the G-study, which provides insights into the various sources of error and their contributions to score variability, the D-study aims to optimize measurement practices and evaluate the reliability of the proposed procedures. (Keiffer, 1998; Huang, 2008). The D-study is essential for determining the adequacy of the measurement procedure for the specific decision-making context as it allows researchers to determine which facets or factors should be prioritized for improvement or control in assessment procedures. (Briesch et al., 2014). Overall, the D-study extends the findings of the G-study by guiding how to improve measurement procedures.

G-theory was employed as the theoretical framework of the quantitative analyses in this study because of its sophisticated and robust nature in the field of ESL/EFL writing assessment. The primary goal was to explore the intricate interplay of several key factors within the assessment process: the number of raters, the variety of tasks presented to the students, and the specific assessment methods employed. In doing so, the study aimed to shed light on how these multifaceted elements collectively influence the variability and reliability of an institutional high-stakes EFL writing assessment procedure.

## 2. METHOD

The present study is a descriptive research as it aims at describing the existing situation without manipulating the variables and making the necessary determinations based on the data obtained. This descriptive study incorporated both quantitative and qualitative data to answer the research questions. The quantitative data were collected to find out the variability and reliability of scores obtained from this specific assessment procedure while the qualitative data were collected to search out the raters' perspectives of the scoring procedure.

### 2.1. Selection of Writing Samples

The writing samples of this study were collected from the School of Foreign Languages at a Turkish state university in the 2022-2023 academic year. Forty-five B1-level students (19 female and 26 male, aged 18 to 24) from the English preparatory program were required to write two essays in separate sessions, as it is impossible to assess task effects using a single-task scenario within the G-theory framework. In the first session, the students were required to write a narrative essay on *"Write about your worst, best, or most embarrassing time in your life"*. In the second session, they were tasked to write an opinion essay on *"Write about advantages and disadvantages of living in a big city"*. The topics were selected from the institutional English proficiency exams administered in the previous years. Following the same procedure administered in the institutional exams, the students were required to write each of their 200-220 word essays in 30 minutes using pen and paper. Totally 90 essays were collected from the students. Then, to ensure a wide range of variation among the essays, two independent raters, who did not participate as raters in the scoring procedure of the study, meticulously categorized the essays into three qualities (i.e., high, medium, and low) using the holistic scoring scale used in the scoring procedure. The raters did not assign numerical scores to the essays during this process. Only the essays that were consistently classified as having either high- or low-quality by both raters were selected for further analysis. As a result, a total of 32 essays, written by 16 students, were determined to be used as the sample for the current study.

### 2.2. Selection of Raters

The purposive convenience sampling method was used to select the EFL instructors based on their willingness to volunteer their time and their proximity to the researcher (Creswell, 2012). The raters had to meet the following criteria: a) being a full-time employee at an EFL teaching institution, b) having experience in teaching EFL writing, and c) having participated in the institutional high-stakes writing assessment. As a result, ten instructors, consisting of six females and four males, took part in this study as raters. They were highly skilled in EFL teaching, boasting expertise in teaching and assessing writing with at least ten years of experience. The instructors were full-time employees of a Turkish state university and native Turkish speakers, with ages ranging between 36 and 52 with a mean of 43. All of the raters were informed about the purpose of the study and they wholeheartedly agreed to participate in the study. To ensure privacy and confidentiality, the participants' identities were kept confidential through the use of pseudonyms.

### 2.3. Scoring Rubrics

One of the primary objectives of this study is to investigate how the choice of scoring method impacts the variability and reliability of scores. To achieve this, the raters were tasked with evaluating the essays twice, employing two different approaches: initially utilizing a holistic method, followed by an analytical approach, with a three-week time interval. The holistic scale was the authentic institutional scale used for the high-stakes writing assessment, which required the raters to assign a single overall score, out of 100 points, to an essay based on its content and organization, language use, and mechanics. An adapted version of the analytic scale Jacobs et al. (1981) developed was used in analytic scoring because its scoring criteria were compatible with those of the holistic scale, but this time they were required to assign a score for each of the five categories: a) content (30 pts.), b) organization (20 pts.), c) grammar (20 pts.), d) vocabulary (20 pts.), and e) mechanics (10 pts.).

### 2.4. Scoring Procedure

Before the scoring procedure, the raters were thoroughly informed of the purpose of the study and presented with a consent form ensuring the protection of their rights and the confidentiality of the obtained data. Following this, the raters were introduced to the holistic scale, and they assessed three essays representing different proficiency levels (low, medium, and high) to build

a common understanding of the scoring criteria they used. They discussed the differences in their scores to align their expectations and judgments. Then, the raters were given a set of materials, which included 32 essays on two different topics, one holistic scoring rubric, one scoring form to write the scores on, and a questionnaire that was formed to gather background information about the raters. Three weeks after they completed holistic scoring, they were introduced to the analytic scale. The three-week time interval was set to prevent paper familiarity. The components of each level on the scale and what they signified were explained until the expectations were all clear. Once again, the raters evaluated three essays representing varying proficiency levels analytically and discussed the disparities in their scores. Finally, the raters were required to score the 32 essays analytically. The raters did not receive extensive training for holistic and analytic scoring in this study, as they had already been trained in assessing institutional exam papers.

### 2.5. Interviews with Raters

After completing both the holistic and analytic scoring procedures, all raters were interviewed individually to gather their perceptions of the single-task, single-rater, and holistic scoring methods used in their institution. Each interview lasted nearly 15 minutes with four main questions regarding the number of writing tasks, the number of raters, the scoring method, and the current assessment procedure in general. Some extra questions were asked when it was felt necessary to get further explanation on the answers. The interviews were carried out in Turkish to gather more detailed information. The interviews were recorded, transcribed, and then translated into English by the author of this study, which were checked by another researcher who had experience in analysing qualitative data.

### 2.6. Data Analysis

This study utilized the G-theory framework to analyze quantitative data to investigate the influence of various factors such as paper, task, rater, and their interactions on the variance of scores obtained from holistic and analytic scoring using the EduG computer program. The researcher conducted two distinct G-studies, one dedicated to holistic scoring and the other to analytic scoring. Each of these G-studies took into account the random effects of the combination of individuals, tasks, and raters, denoted as person-by-task-by-rater (p×t×r). By separately analyzing holistic and analytic scoring, the study aimed to gain a nuanced understanding of how these different approaches contribute to score variance, shedding light on their specific strengths and weaknesses. Furthermore, the research delved into a separate realm of analysis through two random effects D-studies, one for each scoring method (holistic and analytic). These D-studies were conducted to calculate generalizability coefficients, which are typically used in norm-referenced tests to assess the extent to which assessment outcomes can be generalized, and dependability coefficients, which are employed in criterion-referenced tests to gauge the reliability of the assessment process. The D-studies were executed with varying numbers of raters and tasks, offering insights into the impact of these key variables on the reliability and validity of the scoring methods. The culmination of these analyses not only enriched our understanding of the assessment processes but also furnished valuable insights for future test design and evaluation practices.

Furthermore, the qualitative data obtained through the rater interviews were analysed through manual content analysis as suggested by Creswell (2012). The author of this study compiled the student answers under each interview question. The author proceeded to conduct a more in-depth examination of the compiled student answers. The data were carefully scrutinized, and similar responses were grouped together under specific categories. This process was carried out by both the author and another experienced researcher, who worked independently to ensure that their categorization was unbiased. Then, the author and the researcher worked together to sort the categories into themes that corresponded with the interview questions. Direct quotes from the interviews were also included to increase the validity of the qualitative data.

### 2.7. Validity and Reliability of Data Collection Tools and Procedure

To ensure the reliability and validity of both the data collection tools and procedures, several precautions were implemented. First, students generated writing samples under conditions mirroring those of the actual institutional writing exams, with topic selection based on real exam topics tailored to students' proficiency levels and familiarity. Second, two independent raters categorized the collected writing samples into high, medium, and low qualities and the papers which the two raters agreed to be high-quality or low-quality were selected for data analysis. Third, the raters were introduced to the criteria of holistic and analytic rubrics before the scoring procedure. They individually scored three sample essays using these rubrics and engaged in discussions until a consensus was reached on their understanding of the criteria and expectations. This aimed to minimize inconsistencies arising from potential misunderstandings. In addition, a three-week interval was introduced between the holistic and analytic scoring procedures to mitigate rater familiarity with the papers. Finally, to enhance the reliability of qualitative data analysis, the author collaborated with another experienced researcher during the qualitative data analysis procedure.

## 3. RESULTS

### 3.1. The Results of Random Effects Person-by-task-by-rater (p × t × r) G-studies

Specifically, two distinct random effects G-studies, one focusing on holistic scores and the other on analytic scores, were conducted. These G-studies allowed us to scrutinize the multifaceted factors contributing to the overall variance observed in the scoring of the 32 papers. The assessment encompassed a person-by-task-by-rater (p × t × r) framework, which means that we explored how individual students, the specific tasks assigned, and the raters who assessed the papers collectively influenced the final scores. By doing so, we were able to unravel the complex web of interactions among these key components, shedding light on the various aspects that impacted the overall variance in the scoring process. The outcomes of these analyses are given in Table 1.

**Table 1.** *Variance components for random effects p × t × r G-study.*

| Type of Scores | Source of Variability | *df* | $\sigma^2$ | % |
|---|---|---|---|---|
| Holistic Scores | *p* | 15 | .55 | 20.8 |
| | *t* | 1 | .10 | 3.8 |
| | *r* | 9 | .50 | 19.9 |
| | *pt* | 15 | .82 | 30.8 |
| | *pr* | 135 | .10 | 4.1 |
| | *tr* | 9 | .16 | 6.1 |
| | *ptr* | 135 | .65 | 24.6 |
| | *Total* | 319 | 2.63 | 100 |
| Analytic Scores | *p* | 15 | .99 | 38.9 |
| | *t* | 1 | .02 | 0 |
| | *r* | 9 | .26 | 9.8 |
| | *pt* | 15 | .23 | 9.1 |
| | *pr* | 135 | .09 | 3.9 |
| | *tr* | 9 | .04 | 1.7 |
| | *ptr* | 135 | .67 | 26.5 |
| | *Total* | 319 | 2.53 | 100 |

The breakdown of variance components for the holistic scoring, as presented in the Table 1, revealed that the largest contributor to the overall variance was the person-by-task (pt) interaction, accounting for a substantial 30.8% of the total variance. This outcome implies that the 16 EFL students exhibited significantly divergent performance levels in their execution of

the first and second writing tasks. The disparities in their output underscore the distinct challenges posed by these tasks, rendering them non-uniform in nature. Following closely, the residual component (ptr) emerged as the second most influential source of variance, representing 24.6% of the total variance. This component suggests that factors beyond the anticipated interactions among raters, writing tasks, and individual students played a significant role in the variations observed in the scores. These unexplained sources may encompass systematic and random errors, as well as latent factors that eluded detection in the present analysis, thereby underlining the multifaceted and nuanced nature of the holistic scoring process. Person (p) contributed 20.8% of the overall variance, signaling that the evaluation scores assigned to the 16 students were substantially shaped by their characteristics and competencies. These unique traits and skills held a discernible sway over the final scores, reinforcing the idea that the students' inherent abilities were integral to the assessment process. Additionally, the rater component, which represented 19.9% of the total variance, exhibited the raters' varying degrees of leniency or severity in their holistic marking of the papers. In essence, this suggests that the diversity in final scores could be attributed, to a considerable extent, to the idiosyncratic scoring tendencies of the raters. The task-by-rater (tr) component, at 6.1% of the total variance, hinted at the presence of considerable inconsistency among the raters in their evaluation of the two writing tasks. This inconsistency indicates that the raters had differing interpretations of the scoring criteria, further underscoring the intricate nature of the evaluation process. Meanwhile, the person-by-rater (pr) component contributed 4.1% of the total variance, emphasizing that the raters displayed inconsistencies in their evaluation of the essays authored by the 16 EFL learners who participated in this study. This irregularity points to a degree of subjectivity and variation in the raters' judgments. Finally, the task (t) component, representing 3.8% of the total variance, revealed a minor disparity in terms of the difficulty levels of the two tasks. This finding highlights that the tasks were not entirely equivalent in their demands, adding complexity to the holistic scoring process.

The breakdown of analytic scoring components, as outlined in Table 1, showed that the person (p) factor emerged as the most prominent contributor to the total variance, comprising a substantial 38.9%. This observation underscores a crucial point that the analytic scoring approach effectively discriminated among the 16 EFL learners, revealing significant disparities in their respective writing skills. Concurrently, the residual component (ptr), representing unexplained sources of variance, constituted the second-largest share of the total variance at 26.5%. This component serves as a critical reminder that not all aspects of scoring variability can be accounted for, highlighting the inherent complexity of the assessment process. Another salient finding was the rater (r) factor, which accounted for 9.8% of the total variance. This suggests that the raters themselves exhibited discernible differences in their approach, with some demonstrating greater leniency while others leaned towards severity when evaluating the papers analytically. This variance in rater behavior re-emphasizes the importance of consistency among raters in the assessment process. Moreover, the interaction between person and task (pt) contributed to 9.1% of the total variance, indicating that the nature of the writing tasks had a discernible influence on how raters approached analytic scoring. This finding highlights the need to consider the specific writing tasks and their inherent challenges when interpreting the assessment results. The person-by-rater interaction (pr) and task-by-rater interaction (tr) made up 3.9% and 1.7% of the total variance, respectively. These components highlight the complexity of the assessment process, where the interactions between individual learners and raters, as well as between writing tasks and raters, introduce additional layers of variability that can affect the final scores. Interestingly, the task (t) component accounted for 0% of the total variance, indicating that the difficulty of the writing tasks did not influence the raters' analytic scoring. This finding suggests a degree of consistency in the raters' approach across different writing tasks, despite the disparities in individual task complexities.

### 3.2. The Results of Person-by-task-by-rater (p × T × R) Random Effects D-studies

In order to thoroughly examine the reliability of the scores, we conducted two separate D-studies for holistic and analytic scoring, respectively. These D-studies were performed in a person-by-task-by-rater (p × T × R) framework, which means that we took into account variations across different individuals, tasks, and raters. The generalizability coefficient (Ep2) provides insights into the overall consistency and generalizability of the scores, helping us understand how reliably they can be applied in a broader context. The dependability coefficient (ϕ) allowed us to gauge the stability and dependability of the scores within the specific context of our analysis. By conducting these two distinct D-studies for both holistic scoring and analytic scoring, we aimed to understand the reliability and consistency of the scoring methods, which is vital for ensuring the accuracy and validity of our assessment process. The coefficients that are equal to or above 0.70 provide evidence that the scores are consistent and reliable measurements of the writing quality being assessed. The results of the D-studies are presented in Table 2.

**Table 2.** *Generalizability and dependability coefficients.*

| Number of Papers | Number of Tasks | Number of Raters | Holistic Scoring | | Analytic Scoring | |
|---|---|---|---|---|---|---|
| | | | Ep2 | ϕ | Ep2 | ϕ |
| 16 | 1 | 1 | .26 | .21 | .50 | .39 |
| 16 | 1 | 2 | .32 | .27 | .62 | .53 |
| 16 | 1 | 3 | .34 | .30 | .67 | .60 |
| 16 | 1 | 4 | .35 | .31 | .70 | .64 |
| 16 | 1 | 10 | .38 | .35 | .76 | .73 |
| 16 | 2 | 1 | .40 | .31 | .64 | .48 |
| 16 | 2 | 2 | .47 | .39 | .75 | .62 |
| 16 | 2 | 3 | .50 | .43 | .79 | .69 |
| 16 | 2 | 4 | .52 | .46 | .82 | .74 |
| 16 | **2** | **10** | **.55** | **.51** | **.86** | **.82** |
| 16 | 3 | 1 | .48 | .37 | .71 | .52 |
| 16 | 3 | 2 | .56 | .47 | .81 | .66 |
| 16 | 3 | 3 | .59 | .52 | .84 | .73 |
| 16 | 3 | 4 | .61 | .54 | .86 | .77 |
| 16 | 3 | 10 | .64 | .60 | .90 | .86 |

For holistic scoring, as presented in Table 2, the generalizability and dependability coefficients in the current scenario involving 16 essays, two tasks, and ten raters were .55 and .51, respectively. In the single-task, single-rater, and holistic scoring procedure, the generalizability and the dependability coefficients would be .26 and .21, respectively, which would fail to reach the acceptable reliability coefficient of .70. This suggests that relying on a single rater and single task for scoring would result in lower reliability, indicating reduced generalizability of the scores to a larger population. If the number of raters and writing tasks was increased to two in this scenario, the generalizability and the dependability coefficients would be .47 and .39, respectively, which are far below the acceptable reliability coefficient of .70.

For analytic scoring, also given in Table 2, the generalizability and dependability coefficients in the current scenario involving 16 essays, two tasks, and ten raters were .86 and .82, respectively, which are significantly higher than the coefficients obtained from the holistic scoring. If analytic scoring was used in the single-task and single-rater scenario, the generalizability and the dependability coefficients would be .50 and .39, respectively, which are still below the acceptable reliability coefficient of .70 although they are much better than the coefficients obtained from the holistic scoring in the same scenario. If the number of raters and writing tasks was increased to two and analytic scoring was used instead of holistic scoring,

the generalizability and the dependability coefficients would increase to .75 and .62, respectively.

### 3.3. The Findings of the Rater Interviews

To gather the raters' views regarding the overall quality of the current institutional writing assessment procedure, four main questions were asked to the raters in the interviews held after they completed the scoring procedure. The analysis of the data obtained from the rater interviews yielded the following three themes that are related to each interview question: a) using a single writing task is sufficient in assessing students' writing skills; b) using a single rater is not appropriate for high-quality writing assessment; c) analytic scoring method provides more reliable results than holistic scoring method.

First, most of the raters stated that using a single writing task was sufficient in assessing EFL learners' writing skills. Contrary to what is suggested in the literature and what was found as a result of the random effects of person-by-task-by-rater D-studies conducted in the current study, the raters believed that increasing the number of writing tasks would not affect the score reliability. They commented that if the examinees were required to write two tasks, they would get more stressed and tired, which in turn would impact their performance negatively. In addition, they commented that scoring two tasks would not be practical in the high-stakes writing assessment context since a large number of examinees take this test and the results have to be announced in an expeditious manner. Only two of the raters suggested that if the number of writing tasks was increased from one to two, more reliable scores could be achieved.

Second, all of the raters agreed that using a single rater was not appropriate to provide a high-quality writing assessment procedure. They suggested that it is necessary to involve at least two raters in the scoring procedure for reliable and fair results in any high-stakes writing assessment contexts. Regarding this issue, one of the raters reported that *"As raters differ from each other in terms of their scoring behaviours, some raters tend to give high scores while the others tend to give low scores. Therefore, involving two raters in the scoring procedure was effective in decreasing the measurement error stemming from raters' tendencies"*. They also suggested that when the gap between the two raters' scores is large, a third rater should be asked to score the same essay to increase the reliability. In addition, they argued that their scoring performance should be monitored periodically and they should be provided with some feedback regarding their performance. Moreover, they added that the institution should organize more detailed rater training programmes to improve the consistency among the instructor raters.

Finally, it became evident that a significant majority, specifically eight out of the ten raters, agreed that the holistic scoring approach was unsuitable due to concerns regarding score consistency and reliability. They believed that analytic scoring would yield more realistic scores as the rater had to read the essay again and again in order to decide its quality based on the detailed criteria given in the analytic scale. Based on their experiences of scoring the essays for this study, two of the raters made the following comments regarding this issue: *"I could decide the holistic scores after reading the essay only once, but while I was scoring the same essays analytically, I had to read them again to decide the score for each subcategory of the analytic scale (i.e., content, organization, grammar, vocabulary, and mechanics)"*, *"I had to think more about the details regarding organization, grammar, vocabulary, and mechanics while scoring the essays analytically, which made me think that my analytic scores were more accurate than the holistic scores I assigned to the same papers"*. In addition, one of the raters reported that *"I realized that I do not consider mechanics when I score an essay holistically"*. Another rater made the following comment: *"I realized that in holistic scoring the use of language is the component that impacts my score most. If the student can use the grammatical structures accurately, I tend to give a high score even if the content is not sufficient"*. However, another rater stated that *"Content is the most important quality for me while scoring an essay holistically. If the student can explain the topic adequately with necessary supporting details, I*

*do not care about grammatical problems. However, the analytic scale prevented me from ignoring the other components that are necessary for high-quality writing"*. These comments show that the raters demonstrate varying scoring behaviours in holistic scoring, which might increase the variability of scores and thus decrease the score reliability. However, the analytic scale enabled them to consider each subcategory thoroughly while scoring the essays. In addition, the analytic scale limited their overgeneralization of a single aspect of writing. However, a contrasting perspective was voiced by two out of the ten raters who argued that holistic scoring might be a more suitable approach for the high-stakes writing assessment conducted within the institution centering on the belief that holistic scoring proved to be a more time-efficient method as compared to the analytic scoring system.

## 4. DISCUSSION and CONCLUSION

The study utilized G-theory and conducted interviews with raters to explore the influence of a single-task, single-rater, and holistic scoring approach on score variability and reliability within the Turkish higher education context. It was expected that the findings, while specific to this study, could offer valuable insights to assessment experts in various educational institutions. These insights would serve as a blueprint for them to reevaluate and enhance their own writing assessment procedures, particularly in terms of improving score consistency and reliability, extending the potential impact of this research beyond its immediate context.

First, the random effects of person-by-task-by-rater G-studies provided insights into the distribution of variance for the two scoring methods. The results showed that in analytic scoring the variance component attributed to individual persons, defined as the desired variance by Brennan (2001), constituted a substantially larger portion than in holistic scoring. This suggests that analytic scoring was more effective in distinguishing the EFL learners in terms of their writing skills compared to holistic scoring. In the present study, the undesired variance stemming from factors such as the rater, the interaction between individuals and raters, and the tasks and raters (Brennan, 2001) was larger in holistic scoring than it was in analytic scoring. Specifically, the variance attributed to the interaction between the task and raters was over three times greater for holistic scoring than it was for analytic scoring. In line with previous research (e.g., Cumming et al., 2002; Gebril, 2009; Zhao & Huang, 2020), this result indicated that the nature of the task influenced the raters' scores. In the present study, holistic scoring exhibited a greater task effect compared to analytic scoring. Additionally, in holistic scoring the variance associated with the rater accounted for nearly twice as much of the total variance compared to analytic scoring, suggesting that raters exhibited greater inconsistency in their evaluations when employing holistic scoring, particularly in terms of leniency or severity in their ratings. This finding is consistent with prior studies conducted by Barkaoui (2008) and Liu and Huang (2020), but it contradicts the results of Barkaoui's (2010) study, which indicated more rater inconsistency in holistic scoring. Moreover, the variance component referred to as residual, which encompasses the interaction between raters, writing tasks, individuals, and other unexplained systematic and unsystematic sources of error, significantly contributed to score variance in both scoring methods. This underscores the importance of carefully considering and standardizing scoring procedures to minimize measurement errors, as emphasized by Brennan (2001) and Huang et al. (2012).

Second, the person-by-task-by-rater random effects D-studies revealed that the score reliability coefficients obtained from the single-task, single-rater, and holistic scoring procedure would fall significantly short of meeting the acceptable reliability standards for holistic scoring. In contrast, analytic scoring showed more acceptable reliability coefficients. If two writing tasks and two raters were involved in the same assessment procedure, the reliability coefficients would still be lower in the holistic scoring, but in the analytic scoring, the reliability would reach an acceptable level in the norm-referenced assessment while it would be lower in the criterion-referenced assessment. These results revealed that, in accordance with existing

research (Lee et al., 2002; Liu & Huang, 2020; Zhao & Huang, 2020), increasing the number of raters and writing tasks would have a positive impact on the reliability coefficients in both holistic and analytic scoring methods. However, it's important to note that even with these improvements, holistic scoring would not reach satisfactory reliability coefficients. On the other hand, by opting for analytic scoring and concurrently increasing the number of raters and tasks, the assessment process would have a significant enhancement in terms of score reliability. In summary, the results suggest that while holistic scoring benefits from more raters and tasks, switching to analytic scoring would result in notably improved score reliability.

Finally, the findings obtained from the rater interviews showed that the raters were mostly positive about using a single-task in the high-stakes writing assessment procedure because they thought it was more practical and time-efficient in such an assessment context where a large number of examinees' papers must be scored in a short time. In addition, contrary to what the literature suggested and the quantitative results of this study showed, they believed that a single writing task would be sufficient to measure the EFL learners' writing performance. On the other hand, in line with what the literature suggested (Gebril, 2009; Weigle, 2002), the raters did not favour using a single rater in the assessment of high-stakes writing tests as it would endanger the reliability and fairness of the scores. They believed that involving two raters in the scoring procedure can provide more reliable scores. Further, the raters were mostly positive about the analytic scoring method giving the reason that it would yield more realistic and reliable scores because when scoring the essays analytically, they were to abide by the criteria specified in the scale rather than making decisions based on their personal judgments, as supported by the related literature (Barkaoui, 2008; Barkaoui, 2010). Further, in line with the literature (Attali, 2020; Fahim & Bijani, 2011; Weigle, 1994), they commented that receiving rater training periodically might alleviate the inconsistencies stemming from different rater behaviours.

Overall, the results of this study demonstrated that the single-task, single-rater, and holistic scoring procedure would not be sufficient to guarantee high-quality in terms of reliability and fairness issues. Since writing scores are used for making important decisions about examinees in Turkish higher education, it is crucial to make some revisions in the single-task, single-rater, and holistic scoring procedure in order to ensure low variability and high reliability of scores. For this reason, in light of the findings of this study, it is suggested that examinees are required to write at least two writing tasks, and these tasks are scored by at least two raters employing the analytic scoring method. Including a third rater in the scoring procedure when the gap between the two raters is large, might also be a solution to increase the score reliability. In addition, instructor raters must be provided with training for the implementation of the revised assessment procedure. They should be monitored at regular intervals and given feedback about their scoring performance. The assessment policy makers in the Turkish higher education context should consider these suggestions while designing the EFL writing assessment procedures to attain sound and reliable results and make appropriate improvements in EFL education provided in Turkish higher education. Following these suggestions can guarantee the quality of high-stakes writing assessment procedures.

It's essential to recognize two limitations of this study when interpreting its results. Firstly, the study was not carried out in a real high-stakes writing assessment environment, meaning that the data collected may not precisely mirror what occurs in an authentic setting. Raters and examinees might respond differently under the pressures and conditions of a genuine test. Secondly, the relatively small number of selected papers used in this study could restrict the generalizability of the findings to a broader context. To enhance the generalizability of these findings, future research should encompass a broader selection of papers and diverse EFL writing assessment scenarios within Turkish higher education. This will enable a more comprehensive understanding of the factors in different contexts.

## Acknowledgments

I would like to express my gratitude to my colleagues, who wholeheartedly participated in this study as raters. Thank you all for your invaluable time.

## Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number:** Karadeniz Technical University, Ethics Committee for Social and Human Sciences, E-82554930-050.01.04-421870.

## Orcid

Elif Sarı https://orcid.org/0000-0002-3597-7212

## REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Attali, Y. (2020). Effect of Immediate Elaborated Feedback on Rater Accuracy. *ETS Research Report Series*, *2020*(1), 1-15.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* [Unpublished doctoral dissertation, University of Toronto, Canada].

Barkaoui, K. (2010). Do ESL essays raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31-57.

Brennan, R.L. (2001). *Generalizability theory: Statistics for social science and public policy*. Springer-Verlag. Retrieved from https://www.google.com.tr/search?hl=tr&tbo=p&tbm=bks&q=isbn:0387952829

Briesch, A.M., Swaminathan, H., Welsh, M., & Chafouleas, S.M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of Psychology, 52*(1), 13-15. http://dx.doi.org/10.1016/j.jsp.2013.11.008

Cheong, S.H. (2012). Native-and nonnative-English-speaking raters' assessment behavior in the evaluation of NEAT essay writing samples. *영어교육연구*, *24*(2), 49-73.

Creswell, John W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4 th ed.). Pearson Education.

Cronbach, L.J., Gleser, G.C., Nada, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67-96.

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communica tion Quarterly, 1*(1), 2-24.

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.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all?. *Language Testing*, *26*(4), 507-531.

Güler, N., Uyanık, G.K., & Teker, G.T. (2012). *Genellenebilirlik kuramı*. Pegem Akademi Yayınları.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing* (pp. 69-87). United Kingdom: Cambridge University Press. https://doi.org/10.1017/CBO9781139524551.009

Hamp-Lyons, L., & Mathias, S.P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49-68. https://doi.org/10.1016/1060-3743(94)90005-1

Han, T., & Huang, J. (2017). examining the ımpact of scoring methods on the ınstitutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, *53*, 112-147.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? - A generalizability theory approach. *Assessing Writing*, *13*(3), 201-218. http://dx.doi.org/10.1016/j.asw.2008.10.002

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal*, *2*(4), 423-443. https://doi.org/10.5054/tj.2011.269751

Huang, J., Han, T., Tavano, H., & Hairston, L. (2014). Using generalizability theory toexamine the impact of essay quality on rating variability and reliability of ESOL writing. In J. Huang & T. Han (Eds.), *Empirical quantitative research in social sciences: Examining significant differences and relationships*, (pp. 127-149). Untested Ideas Research Center.

Huot, B.A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*, 201-213. https://www.jstor.org/stable/358160

Huot, B. (2002). *(Re)Articulating writing assessment: Writing assessment for teaching and learning.* Logan, Utah: Utah State University Press.

Jacobs, H.J., Zingraf, S.A., Wormuth, D.R., Ha rtfiel, V.F., &Hughey, J.B.(1981). Testing ESL composition: A practical approach. Massachusetts: Newbury House.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* The Guilford Press.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177-182.

Kieffer, K.M. (1998). *Why generalizability theory is essential and classical test theory is often inadequate?* Paper presented at the Annual Meeting of the South Western Psychological Association, New Orleans, LA.

Kenyon, D. (1992, February). *Introductory remarks at symposium on development and use of rating scales in language testing*. Paper presented at the 14th Language Testing Research Colloquium, Vancouver, British Columbia.

Kim, A.Y., & Gennaro, D.K. (2012). Scoring behavior of native vs. non-native speaker raters of writing exams. *Language Research*, *48*(2), 319-342.

Lee, Y.-W., Kantor, R., & Mollaun, P. (2002). Score dependability of the writing and speaking sections of new TOEFL. [Proceeding]. *Paper Presented at the Annual Meeting of National Council on Measurement in Education*, New Orleans: LA. Abstract retrieved on December 11, 2012 from ERIC. (*ERIC No. ED464962)*

Liu, Y., & Huang, J. (2020). The quality assurance of a national English writing assessment: Policy implications for quality improvement. *Studies in Educational Evaluation*, *67*, 100941.

McNamara, T.F. (1996). *Measuring second language performance.* Addison Wesley Longman.

Popham, J.W. (1981). *Modern educational measurement.* Englewood: Prentice.

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal*, *85*, 189-209.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A premier.* Sage

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303-325. https://doi.org/10.1177/026553220101800303

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-Speaking, and ESL students?. *Journal of Second Language Writing*, *5*, 163-182.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors?. *Language Testing*, *37*(3), 311-332.Weigle, S.C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223. http://dx.doi.org/10.1177/026553229401100206

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. Assessing Writing, *6*(2), 145-178.

Weigle, S.C. (2002). *Assessing writing*. United Kingdom: Cambridge University Press.

Weigle, S.C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL writing: A pilot study. *TESOL Quarterly*, *37*(2), 345-354.

Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation*, *67*, 100911.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 675–698
https://doi.org/10.21449/ijate.1502786
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Self-regulated learning support in technology enhanced learning environments: A reliability analysis of the SRL-S rubric

**Slaviša Radović**[1*], **Niels Seidel**[1]

[1]Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA), FernUniversität in Hagen, Germany

**ARTICLE HISTORY**

*Received: June 20, 2024*

*Accepted: Sep. 14, 2024*

**Keywords:**

Self-regulated learning,
SRL-S rubric,
Validity,
Reliability.

**Abstract:** Advanced learning technologies have become a focal point in recent educational research, holding the promise of enhancing students' self-regulated learning (SRL) by facilitating various processes of planning, monitoring, performing, and reflecting upon learning experiences. However, concerns have arisen regarding the efficacy and design of technologies, the spectrum of possibilities for SRL support, and too ambiguous claims associated with these technologies. To address these uncertainties and to provide a platform for generating the more empirical evidence, Self-Regulated Learning Support (SRL-S) rubric was developed to facilitate the assessment of SRL support in technology-enhanced learning environments. It is grounded in established educational theory and proven empirical research results. This article presents a study that extends the application of the rubric to establish its reliability and validity, filling a gap in prior research. First, content, criterion-related, and construct validation were performed through international and interdisciplinary experts' reviews. Subsequently, inter-rater and intra-rater reliability were assessed using Intraclass Correlation Coefficients and Cohens Kappa tests. The outcomes of these analysis demonstrated that the SRL-S is a reliable and valid instrument for assessing the levels of SRL support within learning environments. Additional implications for further research to support self-regulated learning are discussed.

## 1. INTRODUCTION

Over the last two decades, there has been a substantial advance in offering online and distance learning environments within higher education (Ameloot et al., 2024). This trend can be attributed to several factors, including the evolving demands of the labor market, the increasing importance of lifelong learning, and the innate desire of individuals to acquire knowledge (OECD, 2019; Mirriahi et al., 2018). Consequently, numerous higher education institutions have taken proactive steps to organize learning materials and offer educational opportunities tailored to diverse groups of students, thereby ensuring the provision of inclusive and high-quality education for all (Wu et al., 2023).

These modern distance and online learning environments (LE) exhibit a range of distinctive advantages. For example, a notable benefit is the flexibility they afford students, granting them the freedom to choose when, what, and where they learn. Additionally, these environments

*CONTACT: Slaviša RADOVIĆ ✉ slavisa.radovic@fernuni-hagen.de FernUniversität in Hagen, Universitätsstr. 11 – IZ / Building 3, Room 2G 13 (2nd floor), 58097 Hagen, Germany



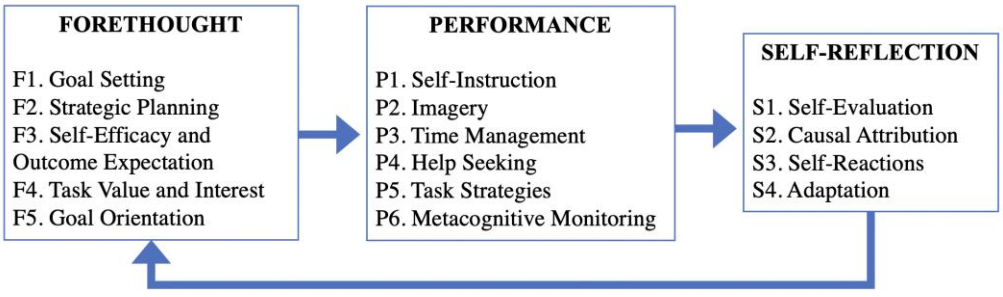
**e-ISSN: 2148-7456**

attract a diverse array of students, each possessing varying levels of prior knowledge, professional experience, and expertise (Mirriahi et al., 2018). Furthermore, they use a specific strategy that requires less direct guidance from instructors (Zimmerman, 2008), fostering greater autonomy among students and providing convenient access to a wide spectrum of learning resources. Despite the apparent benefits, its effectiveness can vary among students including high dropout rates, procrastination, and the long study duration (Goda et al., 2022). While some excel, others may face challenges (Wu et al., 2023).

Empirical research has shown that the acquisition of self-regulated learning (SRL) skills has assumed a critical role in fostering effective and efficient learning (Jivet et al., 2017; Sghir et al., 2022). SRL encompasses a multifaceted set of strategies and learning processes that encompass goal setting, continual progress monitoring, adaptive behavioral adjustments, comprehensive outcome assessment, and reflection (Wu et al., 2023). Students who proactively take control of their own learning processes tend to experience a wide array of academic and non-academic advantages when compared to their peers who are less self-regulated. Nevertheless, many students encounter difficulties when it comes to self-regulation practices. They often struggle with reflective thinking and face challenges in effectively monitoring their progress in alignment with their learning objectives (Radović et al., 2024b). This issue has received significant attention and recognition in academic literature.

From an academic standpoint, SRL has been a widely examined theoretical construct that delineates the cognitive, motivational, and behavioral strategies employed by learners to oversee and govern their own learning processes and results (Zimmerman, 2008; Lodge et al., 2019; Pintrich, 2000). Among the influential models within this domain is Zimmerman's SRL model, which drew upon the foundational work of Bandura and Pintrich. Zimmerman's model articulates three distinct phases in the SRL process: firstly, the thought phase, during which learners set objectives, gauge their motivation levels, and engage in task analysis processes like goal establishment and strategic planning; secondly, the performing phase, wherein learners concentrate their attention, actively participate in tasks, and continually monitor their progress; and lastly, the self-reflection phase, where learners critically assess both the task at hand and their own performance, culminating in comprehensive self-evaluation and self-assessment (Zimmerman, 2008). The complexity of the SRL process and the necessity of aiding students in developing these essential skills has become a paramount concern in both practical educational settings and academic discourse (Wu et al., 2023).

**Figure 1.** *The phases of self-regulated learning, as introduced in Zimmerman (2000) model, with corresponding learning processes and strategies (Radović & Seidel, 2024a; 2024b).*




In light of previous concerns, the remainder of our paper is structured as follows: Section 2 first delves deeper into a range of advanced learning technologies used to effectively and efficiently support students' SRL in distance and online higher education learning environments. Here the focus will be particularly on those technologies based on learning analytics and data mining. The section will then explain the challenging aspect of the SRL support reflecting possible spectrum of variability. Section 3 outlines the research questions addressed in this study, while Section 4 details the research methodology used for data collection and analysis. In Section 5,

we present our findings and engage in a comprehensive discussion of the results. Finally, the article concludes by considering its limitations and offering directions for future research.

## 2. SRL SUPPORT IN LEARNING ENVIRONMENTS

In light of the growing significance of the SRL concept, which, owing to its intricacies, presents a multifaceted challenge, the endeavor to aid students in cultivating these skills remains a central issue for educators and researchers worldwide (Andrade & Du, 2007; Lodge et al., 2019; Mirriahi et al., 2018; Radović et al., 2024a). Empirical research has unequivocally demonstrated that when supported, learners can make substantial progress in enhancing their ability to strategize, monitor, and assess their own learning processes (Ameloot et al., 2024; Goda et al., 2022).

Therefore, various frameworks and advanced learning technologies have emerged in this pursuit, including personalized education, intelligent tutoring systems, adaptive learning systems (Wu et al., 2023; Wang et al., 2023). Insightful review studies conducted by Molenaar et al. (2023), Jivet et al. (2017), Sghir et al., (2022) and other scholars have illuminated a set of specific technological features within learning environments that have proven to be highly effective. These encompass the integration of learning analytics dashboards, provision of support for goal setting, incorporation of self-assessment features, facilitation of guidance for student reflection, and the implementation of personalized recommendations. Refer to Table 1 for a brief overview, and consult the comprehensive review provided by Radović and Seidel (2024a).

**Table 1.** *Advanced learning technologies within learning environments that have proven to be effective for self-regulated learning support.*

| Feature | Description |
|---|---|
| Learning analytics dashboards (LAD) | Learning analytics and data mining techniques can be effectively utilized to develop learning analytics dashboards, as demonstrated by Jivet et al. (2017) and Radović et al. (2024b). These dashboards provide visual summaries of various learning metrics, encompassing factors such as correct and incorrect response rates, time allocation for activities, overall progress, and behavioral patterns (Ameloot et al., 2024; Dong et al., 2024). These metrics can be personalized and adapted to the learner, the learning process, and the learning context. Integrating such features into educational settings empowers students to actively monitor and manage their own learning experiences, as highlighted by Wang et al. (2023). Students can align their efforts with personalized learning plans, assess their progress, and make necessary adjustments for similar tasks in the future, as suggested by Jivet et al. (2017). |
| Goal setting support | Recent comprehensive reviews conducted by Dong et al. (2024) and Jivet et al. (2017) underscore the critical importance of students' ability to select and adapt goal orientations throughout their learning journey. In educational environments, it is essential to design tools and features that assist learners in explicitly defining goals and benchmarks for their learning activities within the curriculum. These support for goal setting should encompass a wide array of performance indicators, progress markers, effort allocation, and criteria for success. It's crucial that these tools effectively integrate the diverse range of learning materials available, including readings, tasks, and self-assessment activities (Radović et al., 2024b). For students, the process of choosing and establishing goals serves two fundamental purposes. Primarily, it offers them guidance and a sense of purpose, influencing their planning and shaping their future actions (Sghir et al., 2022). Secondly, it empowers them to monitor their progress, assess the efficacy of their strategies, and make necessary adjustments to ensure the attainment of their goals. |

| | |
|---|---|
| Reflection support | Reflection is a pivotal component of Self-Regulated Learning (SRL), as briefly noted earlier (Panadero, 2017). It's a cognitive and emotional process, through which learners critically assess their progress, effort, and adapt their learning strategies (Andrade & Du, 2007; Radović, 2024). While reflection is complex and demands initiation, time, and effort, instructions and guiding questions can assist learners in developing reflective thinking skills and becoming more adept at reflective practice (Jivet et al., 2017). Furthermore, directing reflective thinking towards specific learning goals or potential challenges can help learners maintain focus and avoid irrelevant exploration (Zimmerman, 2008). |
| Self-assessment support | Self-assessment is a crucial strategy in higher education, empowering students to independently evaluate their understanding and proficiency in a subject (Andrade & Du, 2007; Panadero et al., 2016). It promotes self-regulated learning by increasing awareness of the learning process and individual responsibility - students review their work, identify performance gaps, and assess against predefined criteria. Additionally, analyzing students' performance and progress in relation to their chosen learning goals, could additionally provide valuable feedback, empowering students to adjust their learning strategies accordingly (Radović et al., 2024a; Wang et al., 2023). |
| Practical recommendations | Adaptive and personalized learning environments are designed to assist learners by tailoring content to their specific needs (Wang et al., 2023). Visual cues can aid learners in adjusting their plans to achieve their goals, but these recommendations are meant to complement, not replace, the SRL process (Ameloot et al., 2024). This is especially valuable for students who face difficulties in self-regulated learning or need additional guidance (Dong et al., 2024 ). This supplementary support can be particularly beneficial for students who may face challenges in practicing SRL, lack clear direction in their learning, experience disorientation or cognitive overload when pursuing their goals, or struggle to identify alternative strategies and strategically plan their learning (Lodge et al., 2019; Radović et al., 2024b). Adaptive and personalized learning environments aim to help learners navigate the complexity of their educational journey by tailoring content to their specific needs at any given moment (Wang et al., 2023). |

### 2.1. Spectrum of SRL Support

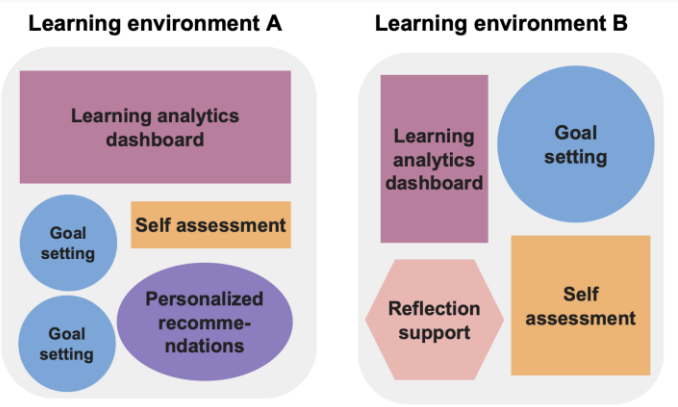
It is widely acknowledged that in order to effectively guide learners through all phases of the SRL cycle, a learning environment must provide a comprehensive and cohesive array of technological features (Radović et al., 2024b). Nevertheless, previous research efforts have often narrowly focused on specific aspects of support. For instance, some studies have concentrated on implementing learning dashboards or only incorporating self-assessment tasks (Pérez-Álvarez et al., 2018; Jivet et al., 2017). Additionally, literature reviews have highlighted an uneven emphasis on different phases of the SRL process, with certain learning environments claiming to support SRL by emphasizing self-monitoring but overlooking self-reflection phase, or vice versa (Goda et al., 2022; Heikkinen et al., 2022).

It has also become evident that SRL support is not a binary concept but rather exists along a spectrum. A recent empirical study conducted by Radović et al (2024b), comparing two learning environments with differing levels of SRL support, revealed that depending on technological features, the levels of SRL support can range from limited to advanced. The results of this study acknowledge that different levels of SRL support can differentially affect students' learning progress and outcome (Radović et al., 2024b). Another research study conducted by Goda et al. (2022) delved into the effects of two learning environments. Case 1 involved an early warning system predicting potential student dropouts, while Case 2 focused on student planning and implementation phases within the self-regulated learning cycle. Their

comparison revealed distinct differences, highlighting that an early warning system requiring pre-learning planning could reduce the necessity for teacher intervention, decrease procrastination tendencies, and result in heightened learning outcomes.

Discrepancies may arise in the developmental scope and feature availability of educational settings, as highlighted by Sghir et al. (2022). Consequently, these variations can influence the level of support they offer for self-regulated learning, as visually depicted in Figure 2 and discussed by Radović and Seidel (2024a). Let's consider Learning Environments A and B, which share identical curriculum content and employ similar technologies. Despite these similarities, the divergence in their support for self-regulated learning becomes evident (Radović and Seidel, 2024a; 2024b; Radović et al., 2024b). Although both environments incorporate sophisticated learning technologies to enhance students' self-regulation, differences in their implementation methods and extents may lead to varying levels of support for self-regulated learning. However, the extent to which these distinctions between the two learning environments are substantial, relative, or absolute, and their potential impact on disparate learning outcomes and processes, remains unverified in the existing research literature. This variability in self-regulated learning support within learning environments poses a significant challenge for researchers and educators, complicating efforts to comprehensively understand and compare diverse developments in this field (Radović et al., 2024b).

**Figure 2.** *Simplified example of difference between two learning environments.*




## 2.2. Rubric for Evaluating the Spectrum of SRL Support

To bear with this challenging aspect of the spectrum of SRL support, Radovic and Seidel (2024a) introduced the rubric, designed to assess the degree of self-regulated learning support available within technology enhanced learning environments (Figure 3 and Appendix A). It is strongly grounded in the theoretical Zimmerman's model (Panadero, 2017) and empirical results distilled from review studies (e.g. Jivet et al., 2017; Pérez-Álvarez et al., 2018; Viberg et al., 2020) that have demonstrated significant effectiveness in supporting student self-regulation. Rubric development process included several phases that will be disclosed in the following text.

**First,** the structure of the SRL-S rubric was developed by mapping the phases of Zimmerman's SRL model (Forethought, Performance, and Self-Reflection) to the dimensions of the rubric (with same titles). Each phase of Zimmerman's model contains multiple learning strategies; for example, the Forethought phase includes Goal Setting, Strategic Planning, Self-Efficacy, Task Value and Interest, and Goal Orientation. These strategies were incorporated as items in the SRL-S rubric (for the corresponding dimension). Therefore, following the SRL model (see Figure 1), our rubric consists of 14 items across the three dimensions: Forethought (F1. Goal Setting, F2. Strategic Planning, F3. Self-Efficacy, F4. Task Value and Interest, F5. Goal Orientation), Performance (P1. Self-Instruction, P2. Imaginary, P3. Time management, P4. Help Seeking, P5. Task Strategies, P6. Metacognitive monitoring), and Self-Reflection (S1.

Self-Evaluation, S2. Casual attribution, S3. Self-reactions, S4. Adaptation). Additionally, each of the items (learning strategies) has been supplemented with a brief description based on Zimmerman's theoretical model (see Table 2).

**Second,** we aimed to gather and analyze review studies that systematically examine the features of advanced learning technologies. Using a broad search strategy, we collected ten systematic reviews of empirical studies focused on tools that support SRL (Araka et al., 2020; Ceron et al., 2021; Devolder et al., 2012; Edisherashvili et al., 2022; Gambo & Shakir, 2021; Garcia et al., 2018; Jivet et al., 2017; Matcha et al., 2020; Pérez-Álvarez et al., 2018; Viberg et al., 2020). We examined how each technology facilitated critical aspects of SRL as outlined in the reviews, considering established clear and distinct standards for each criterion. Each feature and tool are referenced with the review study from which it originated (see the Table 2's column of practical aspects of the rubric). The first author conducted a thorough review of all the studies, identifying key features and tools and categorizing them accordingly. To quantify inter-rater agreement, the second author independently reviewed three recent studies (Ceron et al., 2021; Edisherashvili et al., 2022; Gambo & Shakir, 2021) and categorized the data. Cohen's κ was calculated to assess the level of agreement, showing agreement between the researchers' judgments with kappa value of $\kappa = .526$, $p < .001$ (with total percentage agreement of 80%). This result reflects the proportion of agreement beyond chance, and based on Altman's (1999) guidelines, indicate an acceptable moderate strength of agreement.

**Table 2.** *Initial structure and notes for rubric development process.*

| **Theoretical aspect of rubric based on Zimmerman (2000) SRL model.** | | **Practical aspect of rubric based on evidence from review articles examining learning technologies for SRL (see the note for full set of articles)** |
|---|---|---|
| **Phase of SRL** | **Corresponding strategies and its description** | |
| **Forethought Phase** | **F1. Goal Setting**<br><br>*Establishing specific, measurable, and time-bound objectives to provide direction and motivation for learning.* | - Provide possibilities to select or define goals that focus on skill development, performance improvement, or specific learning activities (Gambo & Shakir, 2021; Jivet et al., 2017; Matcha et al., 2020).<br>- Provide mechanisms for setting educational goals and corresponding sub-goals (Ceron et al., 2021; Matcha et al., 2020).<br>- Offer predefined goal hierarchies and clear descriptions to guide students' navigating their learning path (Devolder et al., 2012; Viberg et al., 2020).<br>- Empower students to define their own goals and select relevant indicators (Matcha et al., 2020).<br>- Encourage the practice of setting and revisiting goals and sub-goals during learning process (Edisherashvili et al., 2022; Viberg et al., 2020).<br>- Implement intelligent agents to assist students in choosing and setting goals concerning course content (Edisherashvili et al., 2022).<br>- Supply detailed information on grading criteria and course standards (Matcha et al., 2020). |
| | **F2. Strategic Planning**<br><br>*Developing a structured approach to achieving goals, including planning steps, resources, and timelines.* | - Utilize dashboard visualizations to provide multi-dimensional presentations of student progress, success, and effort (Matcha et al., 2020; Edisherashvili et al., 2022; Gambo & Shakir, 2021; Jivet et al., 2017).<br>- Guide students toward specific activities during their learning process, ensuring alignment with educational goals (Araka et al., 2020).<br>- Support systematic planning through the use of weekly e-journals, supplemented by prompts to encourage ongoing reflection (Edisherashvili et al., 2022).<br>- Implement prompts that encourage planning of learning activities ahead of time, fostering better preparation and time management (Devolder et al., 2012; Edisherashvili et al., 2022).<br>- Send reminders about progress, accompanied by explicit encouragement, to help students stay focused on their learning goals (Edisherashvili et al., 2022; Viberg et al., 2020).<br>- Offer tools (calendar, schedule support, task list) to assist planning the sequence, timing, and completion of activities (Ceron et al., 2021).<br>- Display a visual representation of the learning resources on the main page, making it easily accessible and serving as a constant reference (Edisherashvili et al., 2022).<br>- Provide information on productive learning strategies (Edisherashvili et al., 2022). |

| | |
|---|---|
| **F3. Self-Efficacy and Outcome Expectation**<br><br>*Cultivating a belief in one's ability to succeed (self-efficacy) and expectations of the outcomes of one's efforts to boost motivation and persistence.* | - Utilize dashboard to provide clear and actionable insights into learning progress, success, and effort; helping students identify areas of strength and improvement (Araka et al., 2020; Jivet et al., 2017; Pérez-Álvarez et al., 2018; Viberg et al., 2020;).<br>- Use visualizations (such as radar graphs, line charts, heat maps, mastery grids, cloud tags, and interaction diagrams) to support analysis of learning process (Edisherashvili et al., 2022; Gambo & Shakir, 2021; Jivet et al., 2017; Pérez-Álvarez et al., 2018; Viberg et al., 2020; Matcha et al., 2020).<br>- Send reminders about progress, accompanied by explicit encouragement, to help students stay focused on their learning goals (Edisherashvili et al., 2022; Viberg et al., 2020).<br>- Provide opportunities for comprehension checks during and after learning activities, followed by immediate feedback (Edisherashvili et al., 2022).<br>- Compare learners' performance with peers who have similar goals, previous graduates, top-performing peers, or teammates (Jivet et al., 2017; Pérez-Álvarez et al., 2018).<br>- Use goals standards to describe outcomes of one's effort during learning (Jivet et al., 2017).<br>- Predict student performance, enabling timely interventions and personalized feedback (Araka et al., 2020; Viberg et al., 2020). |
| **F4. Task Value and Interest**<br><br>*Identifying and enhancing the intrinsic and extrinsic value of the task to increase engagement and effort.* | - Emphasize the relevance and usefulness of tasks to enhance their engagement with the learning material (Ceron et al., 2021).<br>- Highlight personal significance of tasks and relation to the curriculum to make them more engaging (Ceron et al., 2021).<br>- Prompt learners to activate their prior knowledge, facilitating connections with new material (Edisherashvili et al., 2022; Pérez-Álvarez et al., 2018).<br>- Incorporate example-based learning through the use of real world examples and professional tools (Garcia et al., 2018). |
| **F5. Goal Orientation**<br><br>*Adopting a specific orientation towards goals, such as mastery (learning) or performance (demonstrating ability), to guide learning behavior.* | - Provide students with a predefined goal hierarchy and clear descriptions to help them understand and structure their learning (Devolder et al., 2012).<br>- Use prompts to encourage students stay mindful of their overall learning goals (Edisherashvili et al., 2022; Pérez-Álvarez et al., 2018).<br>- Enable students to define and manage their learning paths by offering customized learning activities (Edisherashvili et al., 2022).<br>- Use different colors to denote various aspects and qualities of learning, helping students quickly identify what need to be improved (Edisherashvili et al., 2022).<br>- Provide features that allow students to analyze their performance against goals, giving them a clearer understanding of their standing (Edisherashvili et al., 2022; Jivet et al., 2017; Matcha et al., 2020; Pérez-Álvarez et al., 2018).<br>- Send personalized feedback to learners to complement their achievements and encourage those who may be falling behind (Edisherashvili et al., 2022). |

| | | |
|---|---|---|
| **Performance Phase** | **P1. Self-Instruction**<br><br>*Using prompts or self-talk to guide one's actions and maintain focus during the task.* | - Provide adaptive support that offer timely feedback to guide learning actions (Araka et al., 2020; Pérez-Álvarez et al., 2018; Viberg et al., 2020).<br>- Ensure that course material is presented in a well-structured manner, utilizing diverse media formats to enhance understanding and engagement (Edisherashvili et al., 2022).<br>- Incorporate self-directed prompts to help learners navigate the platform more effectively, encouraging them to reflect about their learning strategies and actions (Edisherashvili et al., 2022).<br>- Implement automated self-assessments that allow comparison of answers with teacher-prepared solutions (Garcia et al., 2018). |
| | **P2. Imagery**<br><br>*Employing mental visualization techniques to rehearse or envision successful task completion and problem-solving.* | - Facilitate students use of concept-mapping tasks to help them organize and visualize knowledge (Devolder et al., 2012; Pérez-Álvarez et al., 2018).<br>- Incorporate mind-mapping tools that aid in mental visualization (Devolder et al., 2012).<br>- Provide a variety of instructional materials (e.g., watching, discussing, conceptualizing, trying out) and allow learners to choose the modes of instruction and materials (Edisherashvili et al., 2022).<br>- Encourage active learning engagement through tools such as text highlighting, annotation, and summarizing (Edisherashvili et al., 2022; Garcia et al., 2018). |
| | **P3. Time Management**<br><br>*Allocating and managing time effectively to balance task demands and ensure timely completion.* | - Assist students in estimating the time required to complete activities (Ceron et al., 2021).<br>- Display a visual representation of the study plan (course material) on the main page of the learning platform, providing a clear overview of tasks (Edisherashvili et al., 2022; Matcha et al., 2020).<br>- Support learners to analyze their progress relative to their peers and teacher-set expectations, helping them organize time more effectively (Edisherashvili et al., 2022).<br>- Monitor time spent on learning, assessments, and planning, offering insights into how students allocate their time across various activities (Gambo & Shakir, 2021; Jivet et al., 2017; Matcha et al., 2020).<br>- Record the time and reasons for interruptions in study sessions to better understand factors affecting learning (Pérez-Álvarez et al., 2018).<br>- Provide hints and prompts to support time management and enhance learning efficiency (Viberg et al., 2020). |
| | **P4. Help Seeking**<br><br>*Actively seeking assistance or feedback from others when encountering difficulties or needing additional support.* | - Encourage students to seek help from instructors, peers, or external resources when needed (Ceron et al., 2021; Garcia et al., 2018).<br>- Explicitly remind students of the possibility of seeking help during their learning (Edisherashvili et al., 2022).<br>- Facilitate collaboration as a means to improve the learning process through collective input (Edisherashvili et al., 2022).<br>- Promote the exchange of constructive peer feedback in discussion forums (Edisherashvili et al., 2022; Gambo & Shakir, 2021; Garcia et al., 2018; Matcha et al., 2020).<br>- Create an open forum where students can share their thoughts and work-in-progress (Edisherashvili et al., 2022), as well as final product (Edisherashvili et al., 2022; Gambo & Shakir, 2021; Garcia et al., 2018).<br>- Use pedagogical agents to encourage help-seeking, guiding students to resources and support (Gambo & Shakir, 2021). |

| | | |
|---|---|---|
| | | - Incorporate social networks, wikis, blogs, discussion forums or shared learning spaces to facilitate support (Pérez-Álvarez et al., 2018). |
| | **P5. Task Strategies**<br><br>*Applying specific methods or techniques relevant to the task to enhance performance and achieve goals.* | - Advice students in organizing, planning, and managing their study time and tasks, including time allocation, sequencing, and reorganization of instructional materials (Ceron et al., 2021).<br>- Provide criteria and solution to tasks (Edisherashvili et al., 2022; Garcia et al., 2018), as well as hints and feedback to help students understand and correct their errors (Devolder et al., 2012).<br>- Include worked-out examples to illustrate problem-solving methods and concepts (Devolder et al., 2012).<br>- Implement strategies such as sketching (Ceron et al., 2021), mind-mapping, and visualization (Devolder et al., 2012).<br>- Encourage the interpretation, analysis, evaluation, and critical thinking during solving complex problems (Ceron et al., 2021).<br>- Offer guidance on the problem-solving steps students can take (Garcia et al., 2018; Devolder et al., 2012).<br>- Provide hints to students on how to proceed when they encounter errors, (Garcia et al., 2018).<br>- Supply information on effective and efficient learning strategies (Matcha et al., 2020).<br>- Encourage active learning engagement through tools such as text highlighting, annotation, and summarizing (Edisherashvili et al., 2022; Garcia et al., 2018). |
| | **P6. Metacognitive Monitoring**<br><br>*Continuously students one's own cognitive processes, such as understanding and adjusting strategies based on progress and difficulties.* | - Inform students in real time about their knowledge gains, enhancing awareness of their capabilities and progress (Ceron et al., 2021).<br>- Prompt students to assess their understanding (eg. self-assessment task, quizzes, tests) (Edisherashvili et al., 2022; Jivet et al., 2017).<br>- Prompt students to evaluate their behavioral engagement with learning units and different learning materials (Edisherashvili et al., 2022).<br>- Send personalized emails to compliment students on their achievements or encourage those who are falling behind (Edisherashvili et al., 2022).<br>- Process learner activity to provide visual summary, estimate progress, and feedback for improvement (Edisherashvili et al., 2022; Garcia et al., 2018; Jivet et al., 2017; Matcha et al., 2020).<br>- Provide dashboard indicators to help students track their progress towards achieving set goals (Gambo & Shakir, 2021; Garcia et al., 2018; Viberg et al., 2020). |
| **Self-Reflection Phase** | **S1. Self-Evaluation**<br><br>*Reflecting on and assessing the effectiveness of one's performance and strategies in achieving goals.* | - Provide prompts to encourage learners to reflect on their learning experiences (Viberg et al., 2020).<br>- Provide predictions of students' performance to help them gauge their progress (Araka et al., 2020; Jivet et al., 2017).<br>- Provide feedback regarding the productivity and relevance of the learning activities (Edisherashvili et al., 2022; Araka et al., 2020)<br>- Offer opportunities for knowledge tests during and after learning activities (Edisherashvili et al., 2022; Gambo & Shakir, 2021).<br>- Provide a visualization and use of different colors to denote various aspects and qualities of learning process (Edisherashvili et al., 2022; Pérez-Álvarez et al., 2018; Viberg et al., 2020). |

| | |
|---|---|
| | - Analyze students' performance against expectations (eg. standards or class averages) to provide benchmarks for reflection (Edisherashvili et al., 2022; Gambo & Shakir, 2021; Jivet et al., 2017).<br>- Implement a social comparison feature that allows learners to analyze their progress in relation to their peers (Edisherashvili et al., 2022; Jivet et al., 2017). |
| **S2. Causal Attribution**<br><br>*Identifying and analyzing the reasons behind successes or failures to understand the factors influencing performance.* | - Provide information that helps learners assess their ability to complete tasks, enhancing their self-awareness and confidence (Ceron et al., 2021).<br>- Incorporate self-assessment and feedback process to encourage students to examine their misunderstanding (Devolder et al., 2012).<br>- Provide dashboard information on previous learning problems, failures, or challenges (Jivet et al., 2017; Matcha et al., 2020).<br>- Use reflection tasks to support learners in planning, setting goals, and reflecting on their learning processes (Edisherashvili et al., 2022; Viberg et al., 2020).<br>- Provide information about areas needing adaptation (Edisherashvili et al., 2022; Matcha et al., 2020). |
| **S3. Self-Reactions**<br><br>*Evaluating personal reactions to performance outcomes, such as satisfaction, frustration, or motivation, to guide future efforts.* | - Address affective reactions in reflection tasks to help students understand and manage their emotional responses (Ceron et al., 2021).<br>- Provide clear and well-defined expectations for upcoming learning experiences (Edisherashvili et al., 2022).<br>- Increase students' awareness of their emotions by presenting insights from previous learning sessions, which can help them manage their emotional responses (Garcia et al., 2018).<br>- Utilize awareness and dashboard visualizations to address misunderstanding, false expectations, and deactivate negative emotions (Jivet et al., 2017; Matcha et al., 2020). |
| **S4. Adaptation**<br><br>*Adjusting goals, strategies, and approaches based on reflections and evaluations to improve future learning and performance.* | - Provide predictions of students' performance to help them understand their potential outcomes and areas for improvement (Araka et al., 2020).<br>- Enable students to analyze their learning process in relation to goals (Ceron et al., 2021; Viberg et al., 2020).<br>- Incorporate reflection questions and 'look back' prompts to encourage students to think about their future learning (Devolder et al., 2012).<br>- Ask students to reflect on challenges encountered during learning and analyze strategies used or not used to address those challenges (Edisherashvili et al., 2022; Matcha et al., 2020).<br>- Provide feedback (personalized messages) for current problems or suggest goals corrections (Gambo & Shakir, 2021).<br>- Offer information for learning strategies that support learning process (Araka et al., 2020; Viberg et al., 2020). |

*Note: A set of review articles (Araka et al., 2020; Ceron et al., 2021; Devolder et al., 2012; Edisherashvili et al., 2022; Gambo & Shakir, 2021; Garcia et al., 2018; Jivet et al., 2017; Matcha et al., 2020; Pérez-Álvarez et al, 2018; Viberg et al., 2020).*

**Third,** the next step involved setting the rubric's grading criteria into three levels: Limited, Moderate, and Advanced SRL support. For each rubric item, contextualized notes (as shown in Table 2) were organized in three groups to distinctly structure different criteria (Limited, Moderate, and Advanced). Then, we provided description of standards in a more decontextualized manner (see Figure 3 for an example and the full rubric in Appendix A). This decontextualization will allow rubric to be applied across various learning environments, situations, conceptual paradigms, and for different research inquiries. To write these criteria descriptions, we again reviewed theoretical articles by Panadero (2017), Pintrich (2000), and Zimmerman (2000). This iterative process (of theoretical and empirical work) aligns with the recommendations of the National Council on Measurement in Education (NCME) Standards (AERA, 2014).

**Figure 3.** *The part of the SRL-S rubric shows only two SRL criteria (F1 from Forethought and S2 from Self-Reflection phase) with corresponding performance levels.*

| Phase | Process | Limited SRL support (1) | Moderate SRL support (2) | Advanced SRL support (3) |
|---|---|---|---|---|
| Forethought | *F1. Goal Setting* | Students acquire course goals predefined by the teacher, they do not have the option to set or modify their goals within learning environment, nor can they easily access goal related performance indicators. | While students still lack the capability to set or change learning goals themselves in the learning environment itself, however they receive detailed insights about their learning concerning the course's goal. | Students enjoy the flexibility to choose from a range of learning goals (which may include course mastery or just passing) or to set custom goals (content or performance related). Additionally, students are provided with details related to the chosen goal. |
| Self-Reflection | *S2. Causal Attribution* | Students are offered a limited resources to reflect (e.g., knowledge tests and related rubrics). They are not guided nor supported how to reflect on performance or how to evaluate factors of failure. | Students are asked to think about their performance when self-assessing tasks' solutions against criteria. This level of support encourages students to consider the factors that influenced their failures. | Learning environment includes prompted critical reflection tasks after major learning events or learning units. These tasks ask students to think– about their performance, their strengths and weaknesses, as well as to assess their progress toward their goals. |

Finally, in Appendix A, the complete SRL-S rubric, introduced by Radović and Seidel (2024a; 2024b), has been showcased and detailed. By employing the rubric, educators and researchers in charge of a learning environment can 1) gain insights into the extent of implemented SRL approaches, 2) make informed decisions to refine their pedagogical strategies, 3) further develop SRL support of learning environments, and 4) better support students on their journey towards becoming self-regulated learners (Jonsson, A., & Svingby, G. (2007).

## 3. RESEARCH QUESTIONS FOR THIS STUDY

To further substantiate the utility of the SRL-S rubric as an instrument for assessing the level of self-regulated learning support in educational settings, this study aims to establish both reliability and validity. According to the principles of the National Council on Measurement in Education (NCME) Standards (AERA, 2014), reliability and validity analyses are crucial for ensuring that measurement tools are accurate, consistent, and fair. While validity ensures that the tool measures what it is supposed to measure and confirms that it is appropriate and meaningful for the specific context (AERA, 2014, p. 11), reliability refers to the consistency of measurement results over time and across different populations (AERA, 2014, p. 43). These analyses support the ethical and professional use of assessments, guiding effective decision-making and promoting equity in educational and psychological contexts, as emphasized by the NCME standards. Given the absence of such extensive analysis in prior empirical research, it is imperative to ascertain the effectiveness and efficiency of the rubric as a measurement tool (Reddy & Andrade, 2010; Moskal & Leydens, 2000; Thaler et al.,2009).

Hence, the primary research question under investigation in this study is as follows: Does SRL-S rubric demonstrate sufficient reliability and validity for its use to measure self-regulated learning support within online learning environments?

# 4. METHOD

## 4.1. Validity Analysis

According to the standards of American standards (AERA, 2014, p. 11), validity is a critical concept in assessment, referring to the extent to which evidence and theory support the interpretations of scores for their intended purposes. The NCME standards classify different types of evidence that can be used to support the validity of a test. These include Content, Construct, and Criterion-related Validity (AERA, 2014, p. 14, 66, 173).

### 4.1.1. *Participants*

As per the guidelines of the standards, the rubric's validity was assessed through a process of expert judgment (AERA, 2014, p. 25). This ensured that the rubric was both representative of and appropriate for the intended construct (Reddy & Andrade, 2010).

In the first phase, an expert discussion was initiated after the presentation of SRL-S rubric during the scientific meeting of members of CATALPA research center (Center of Advanced Technology for Assisted Learning and Predictive Analytics) of FernUniversität in Hagen in Germany. The group comprised 15 researchers, teachers, and professors who engaged in the use and development of diverse tools aimed at supporting students' self-regulation in research and teaching activities.

In the second phase, feedback on validity of developed rubric was solicited from four distinguished higher education professors, each with extensive research experience and proven excellence in self-regulated learning, learning analytics, and data mining, as evidenced by their numerous academic publications. Our aim was to incorporate interdisciplinary expertise and consider diverse geographic and cultural perspectives (Moskal & Leydens, 2000).

### 4.1.2. *Procedure*

According to the NCME, the experts consulted were asked to make a *Content* assessment (evidence that the rubric content is representative of the domain it's intended to cover and identifies any potential gaps or redundancies), *Construct* assessment (evidence that the rubric accurately measures the theoretical construct it claims to measure), and *Criterion-related* assessment (evidence indicating the extent to which rubric scores correlate with practical development, and the degree to which this is adequately informative) of the developed rubric's criteria and performance levels. Moskal and Leydens (2000) also noted that these are an important aspect of consideration because they examine the extent to which the rubric incorporates the knowledge and technological development of the field that is of interest for a variety of interdisciplinary experts interested in SRL support.

Experts received a set of questions evaluating whether the rubric criteria accurately represent technological development, effectively measure the theoretical construct of SRL, and whether any critical elements are missing (to align with practical development). Additional questions were set for exploring the degree of clarity in the wording, the suitability of the indicator to assess a learning environment, and the relevance of different SRLs levels (e.g. Question 3. Do you clearly understand different levels for each criterion? What was difficult to comprehend? Question 4. Is there a SRL support strategy you consider important that we leave out? To what criteria and performance level it belongs?).

## 4.2. Reliability Analysis

According to NCME standards, reliability refers to the degree of consistency and reproducibility of test results across different times and raters (AERA, 2014). The reliability analysis aimed to ensure that test scores accurately reflect the construct being measured. This involved two key methods: Inter-Rater Reliability (AERA, 2014, p. 44), which measures the consistency of scores assigned by different raters or judges and is crucial for subjective

assessments, and Test-Retest Reliability (AERA, 2014, p. 44), which assesses the stability of test scores over time by administering the same test to the same group on different occasions.

#### 4.2.1. *Participants*

First, four faculty members, comprising researchers who were involved in teaching or researching the same course at a distance university in Germany, independently utilized the rubric to evaluate the level of SRL support their course's digital learning environment provided to students. Second, to analyze consistent scoring across time, two of the researchers were asked to re-evaluate the learning environment two months after the first rating.

Since the evaluators needed to possess a profound understanding of learning material, all details of implemented technological features, and specific pedagogical strategies (for example for goal setting, help seeking, or reflection see Appendix), only teachers and researchers directly involved in the course with profound understanding were being able to make relevant assessment. Expanding the pool of participants was not feasible because individuals unfamiliar with the intricacies of the course would not be able to effectively use the rubric for evaluation purposes. Expanding the number of learning environments used for evaluation was also not feasible because these four evaluators would not be familiar with all the features of the learning environments. More on this later under Limitations and Future Research.

#### 4.2.2. *Procedure*

In this study, we employed a comprehensive approach to assess the reliability of the data generated, utilizing several strategies closely paralleled those utilized in prior research by Harris et al. (2010), Tabachnick and Fidell (2019), and Moskal and Leydens (2019), as well as consistent with NCME standards (AERA, 2014). Because we aimed to include more than two raters, instead of Cohen's kappa coefficient (for two raters) the Intraclass Correlation Coefficient (ICC) was used as the method (Thaler et al., 2009) to compute the interrater reliability of the rubric. This statistical measure, derived from the analysis of variance and based on mean squares representing population variances, has been widely employed to gauge interrater reliability when more than two raters were employed (Tabachnick & Fidell, 2019). In our analysis, the two-way absolute agreement model was applied to compute ICC (McGraw & Wong, 1996).

Additionally, to examine the stability of the rubric's performance over time, we assessed its intra-rater reliability. This involved **first** analyzing the percentage agreement between scores assigned to the same learning environment by the same researchers, two months apart; and **second**, calculating Cohen's kappa (κ) coefficient for these two sets, offering a quantitative measure of the test-retest reliability as suggested in work of Moskal and Leydens (2019).

### 4.3. Learning Environment Used for Rating

The rubric was used to score the course that was specifically designed to foster students' SRL as a component of the completely distance and online bachelor's degree programs in Computer Science at the FernUniversität in Hagen in Germany. During a period of 11 weeks students worked individually, by studying material and doing designed assignments, after which they completed the course by doing the final exam. Specific features were developed to support students' regulation: Dashboard learning overview, Reflection assignments, Self-assessment tasks along with the criteria and feedback, Goal setting feature, and Reading support (Radović et al., 2024a; Radović et al. 2024a).

**Figure 4.** *Dashboard for the learning environment which indicates the progress and performance per type of course material for each course unit including an ultimate reflection task. In the upper left corner, there is a dropdown menu that offers various goals.*


An overview page with a Learner Dashboard served as a collection of all learning resources, such as reading materials and various tasks. These resources were neatly organized by course units in rows, allowing students to easily monitor their progress and access available learning materials with a quick glance (Radović et al. 2024a; 2024b). To enhance student self-regulation, the learning resources were categorized by material type. Furthermore, each learning material was accompanied by two indicators, where applicable: "progress" indicated the extent of completion, while "success" reflected the accuracy or achievement in related activities. To provide personalized support, the learning environment introduced a color-coded scheme. This scheme aimed to align students' progress and success with their individual goals. Green highlighted activities in harmony with the set goal, yellow flagged potential issues, and orange indicated performance inconsistencies (Radović et al., 2024a). The feature for setting goals was presented as a user-friendly drop-down menu just below the Semester overview title (see Figure 3). This allowed students to select from three course goals: Mastery of the content, passing the course, or simply gaining an overview, representing their intention to pursue exams or desired performance. Learning overview dashboard included an additional feature: a reflection prompt located at the end of each course unit (positioned in the fourth column on the right side of Figure 1). This prompt aimed to guide students' reflective thinking toward specific learning objectives or potential learning dilemmas. It assisted students in maintaining focus on their goals, overall satisfaction, and effective learning strategies. Furthermore, self-assessments provided students with supplementary information, including the difficulty level, achieved score, and maximum score, during both the performance and thought phases (Radović et al. 2024b).

## 5. RESULTS AND DISCUSSION

### 5.1. Validity of the SRL-S Rubric

The construct validity of the initial draft of the rubric received in general strong support from comments provided by all expert reviewers. The feedback (total of 40 comments) regarding description of technology integration, the associated levels, and performance indicators, including minor suggestions for different language constructs was thoughtfully considered and integrated into the rubric revision process (Moni et al., 2005; Reddy & Andrade, 2010). According to the NCME standards (AERA, 2014, p. 81), this iterative approach to refinement proved instrumental in better aligning the rubric with intended assessment goals. Expert reviewers also identified few other relevant literature and empirical findings that were thoroughly reviewed and included in the current version of the rubric.

### 5.2. Teachers' Interrater Reliability

The researchers' scores for the SRL-S rubric are reported in the Table 3. This table provides the actual ratings as well as the mean scores and standard deviations for each of the rubric criteria for four raters, for their ratings of the learning environment.

**Table 3.** *The detailed ratings of four raters.*

| SRL | SRL Processes / Strategies | R1 | R2 | R3 | R4 | M | SD |
|---|---|---|---|---|---|---|---|
| Forethought Phase | F1. Goal Setting | 3 | 2 | 3 | 3 | 2.75 | 0.50 |
| | F2. Strategic Planning | 2 | 3 | 2 | 3 | 2.50 | 0.58 |
| | F3. Self-Efficacy and Outcome Expectation | 2 | 2 | 2 | 2 | 2.00 | 0.00 |
| | F4. Task Value and Interest | 2 | 2 | 2 | 2 | 2.00 | 0.00 |
| | F5. Goal Orientation | 3 | 2 | 3 | 2 | 2.50 | 0.58 |
| | Overall Forethought Phase | 2.4 | 2.2 | 2.4 | 2.4 | | |
| Performance Phase | P1. Self-Instruction | 1 | 2 | 1 | 2 | 1.50 | 0.58 |
| | P2. Imagery | 1 | 2 | 1 | 1 | 1.25 | 0.50 |
| | P3. Time Management | 1 | 2 | 1 | 2 | 1.50 | 0.58 |
| | P4. Help Seeking | 2 | 2 | 2 | 2 | 2.00 | 0.00 |
| | P5. Task Strategies | 2 | 2 | 2 | 2 | 2.00 | 0.00 |
| | P6. Metacognitive Monitoring | 2 | 2 | 2 | 3 | 2.25 | 0.50 |
| | Overall Performance Phase | 1.5 | 2 | 1.5 | 2 | | |
| Self-Reflection Phase | S1. Self-Evaluation | 3 | 3 | 3 | 2 | 2.75 | 0.50 |
| | S2. Causal Attribution | 3 | 3 | 3 | 3 | 3.00 | 0.00 |
| | S3. Self-Reactions | 3 | 3 | 3 | 3 | 3.00 | 0.00 |
| | S4. Adaptation | 3 | 2 | 3 | 3 | 2.75 | 0.50 |
| | Overall Reflection Phase | 3 | 2.75 | 3 | 2.75 | | |
| Overall SRL support | | 2.3 | 2.32 | 2.3 | 2.38 | | |

The intraclass correlation coefficient (ICC) was used as the method to compute the interrater reliability of the rubric (Moskal & Leydens, 2019). The ICC estimates and their 95% CI were calculated based on the average measures (k = 4), absolute-agreement, 2-way mixed-effects model (including systematic errors of both raters and random residual errors). The ICC score was .86, 95% CI [.71, .95], suggesting good to excellent interrater reliability between the four raters and their scores on the SRL-S. As a rule of thumb, ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability (Thaler et al., 2009).

Furthermore, the analysis of raters' scores of SRL-S as presented in Table 3, reveals that the raters' overall learning support ranges from a minimum score of 2.3 to a maximum score of 2.38. The results suggest that the raters scored the overall levels of SRL support in the learning environment in a very similar manner (with a margin of differences of only 3.5%).

### 5.3. Teachers' Intra-Ratter Reliability

Intra-ratter reliability involved first analyzing the percentage agreement between scores assigned to the same learning environment, and second examining Kappa coefficient as the extent of agreement between frequencies of two sets of data collected on two different occasions.

To determine percent of absolute agreement, we counted the instances in which raters' first and second ratings for each criterion matched (24 cases) and divided this by the total number of criteria ratings (30). This calculation demonstrates 80% absolute agreement. As a general guideline, suggested by various experts, a percentage of absolute agreement falling within the 70-90% range indicates an acceptable level of agreement (Stemler, 2004). In addition to directly comparing the percent agreement between repeated ratings, we employed Cohen's kappa (κ) test to determine the level of agreement beyond what would be expected by random chance, separately for each of the raters, R1 and R3. An analysis of reliability for the R1 rater revealed

moderate agreement between the ratings (κ = .484, $p$ = .01), while for the R3 rater, an almost perfect agreement between repeated scores was observed (κ = .899, $p$ < .001) (Thaler et al., 2009).

Upon an examination of the scores associated with the ratings of SRL phases, as well as the overall SRL support, a consistent and almost perfect agreement regarding the Forethought Phase and the Reflection Phase becomes evident. Notably, the ratings for the Performance Phase experienced the most significant changes over the time. As a result, this influenced a change in the overall SRL support ratings, shifting from 2.3 to 2.36 and from 2.3 to 2.47. Despite these disparities, the ratings convey the very similar level of SRL support, as depicted in Table 4.

**Table 4.** *Rater scores and the level of absolute agreement between raters evaluating the same learning environment (first and second time), as assessed by two researchers (R1 and R3).*

| SRL | SRL Processes / Strategies | R1 | | R3 | | Agreements |
|---|---|---|---|---|---|---|
| | | First | Second | First | Second | *absolute* |
| Forethought Phase | F1. Goal Setting | 3 | 3 | 3 | 3 | 2/2 |
| | F2. Strategic Planning | 2 | 2 | 2 | 3 | 1/2 |
| | F3. Self-Efficacy and Outcome Expectation | 2 | 2 | 2 | 2 | 2/2 |
| | F4. Task Value and Interest | 2 | 2 | 2 | 2 | 2/2 |
| | F5. Goal Orientation | 3 | 3 | 3 | 2 | 1/2 |
| | Overall Forethought Phase | 2.4 | 2.4 | 2.4 | 2.4 | |
| Performance Phase | P1. Self-Instruction | 1 | 1 | 1 | 1 | 2/2 |
| | P2. Imagery | 1 | 2 | 1 | 2 | 0/2 |
| | P3. Time Management | 1 | 1 | 1 | 2 | 1/2 |
| | P4. Help Seeking | 2 | 2 | 2 | 2 | 2/2 |
| | P5. Task Strategies | 2 | 2 | 2 | 2 | 2/2 |
| | P6. Metacognitive Monitoring | 2 | 2 | 2 | 3 | 1/2 |
| | Overall Performance Phase | 1.5 | 1.67 | 1.5 | 2 | |
| Reflection Phase | S1. Self-Evaluation | 3 | 3 | 3 | 3 | 2/2 |
| | S2. Causal Attribution | 3 | 3 | 3 | 3 | 2/2 |
| | S3. Self-Reactions | 3 | 3 | 3 | 3 | 2/2 |
| | S4. Adaptation | 3 | 3 | 3 | 3 | 2/2 |
| | Overall Reflection Phase | 3 | 3 | 3 | 3 | |
| *Overall SRL support* | | 2.3 | 2.36 | 2.3 | 2.47 | 24/30 |

Several limitations of this study warrant consideration. First, the relatively small number of participants must be acknowledged. Obtaining meaningful assessments from individuals not well-acquainted with the learning environment posed significant challenges. This limitation affected both the inclusion of more diverse learning environments for current participants and the possibility to increase the overall number of participants for the learning environment under consideration. In future, the objectivity could be even further improved by a blind rating or students' rating. However, that may bring new challenges. One of these challenges could be that the knowledge of learning environment is not profound enough, for example a developer of LE would know the features very well, but not their effects on students' learning. Second, our study incorporated exclusively an analysis of a single learning environment. To further increase the reliability of the assessment, a greater variety of LE should be assessed that represent different aspects of SRL including very low to no SRL support. Third, there may be a potential bias in our selection of experts for the validation analysis. Nevertheless, we made efforts to include a highly diverse group of interuniversity, international and interdisciplinary experts with established backgrounds in SRL related research and development practices.

## 6. DISCUSSION and CONCLUSION

With the increasing integration of advanced learning technologies in higher education, it has become evident that support for students' self-regulated learning is not a binary concept. Rather, it encompasses various levels of support. This recognition of diversity presents another challenge for both researchers and educators, complicating the comparison of different

developments, the design of effective pedagogical frameworks, and the determination of the optimal level of self-regulated learning support for specific contexts. In response to this challenge, we have recently developed the Self-Regulated Learning Support (SRL-S) rubric, a tool designed to empirically assess the extent and depth of SRL-S within a learning environment. The purpose of this study was to establish the reliability and validity of the SRL-S rubric. We examined various aspects to determine the consistency of ratings, including intra-rater and inter-rater reliability, and assessed whether the rubric was well-designed in terms of criteria and performance levels to differentiate the various levels of SRL support in educational settings. The results of this study indicate that the SRL-S rubric is both reliable and valid, making it a valuable tool for educators and researchers in higher education.

The validity of the rubric is grounded in the alignment of its criteria and performance levels with the concept it aims to measure. It also takes into account the knowledge and technological developments in the field, which are of interest to a diverse group of interdisciplinary experts focused on SRL support (Jonsson & Svingby, 2007). To ensure its validity, we consulted an international and interdisciplinary panel of experts who conducted qualitative content, construct, and criterion assessments of the rubric criteria and performance levels. Their feedback helped us clarify and refine the rubric's performance levels and align the terminology with the broader research community interested in SRL. Fortunately, no major issues were reported. Regarding the rubric's reliability, we employed interrater and intrarater reliability analyses. Interrater reliability proved to be good, while intrarater reliability demonstrated a moderate to almost perfect agreement between repeated ratings. These findings confirm that the rubric is a reliable instrument, delivering consistent results when used by multiple raters or when used multiple times with some time interval.

Future research endeavors should consider exploring the applicability of the SRS-S across diverse populations beyond Germany and especially in various educational settings, distinct from higher education delivered at a distance. This reliability exploration could expand the scope and utility of this tool. Second, subsequent theoretical and empirical research could further extend the rubric by incorporating students' usage indicators column. Existing research has shown that the mere availability of a technological tool in a learning environment does not guarantee its usage by students (Radović et al., 2024a; 2024b). Moreover, studies have demonstrated that the same learning technology can yield different learning outcomes and lead to different learning processes based on how students employ it (Radović, 2024). Consequently, the SRL-S rubric could serve as a valuable platform for comprehending whether and to what extent students utilize the available SRL support within the learning environment.

Ultimately, the SRL-S rubric can function as an instrument for conducting meta-analyses of literature reviews and empirical studies exploring learning environments published on the topic of SRL. Such research endeavors could contribute significantly to our understanding of optimal SRL support, the relationship between various levels of self-regulation and student success, as well as factors like anxiety, time pressure, and cognitive load. Presently, it is widely believed that more advanced SRL support leads to improved learning outcomes; however, extensive and rigorous empirical evidence to substantiate this claim remains lacking (Jivet et al., 2017). It has also become clear that a one-size-fits-all approach to teaching is inadequate, so finding right levels of SRL support for different educational contexts, educational disciplines, or domain-specific learning processes might also be promising ways for further research (Molenaar et al., 2023). To achieve this aim, this rubric could serve as the missing evaluation method and establish a foundation for better understanding.

## Acknowledgments

This work was funded by the Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA) of the FernUniversität in Hagen. We would like to especially

thank Ioana Jivet, Henrik Bellhäuser, Dennis Menze, Joerg M. Haake, and other colleagues for their feedback and support we have received in the process of rubric making.

## 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

**Slaviša Radović**: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Niels Seidel:** Writing - Review & Editing, Project administration.

## Orcid

Slaviša Radović https://orcid.org/0000-0001-8840-6053
Niels Seidel https://orcid.org/0000-0003-1209-5038

## REFERENCES

Ameloot, E., Rotsaert, T., Ameloot, T., Rienties, B., & Schellens, T. (2024). Supporting students' basic psychological needs and satisfaction in a blended learning environment through learning analytics. *Computers & Education, 209*, 104949.

American Educational Research Association. (2014). Standards for educational and psychological testing (AERA, APA, and NCME). Washington, USA: American Educational Research Association, ISBN 978-0-935302-35-6. Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159–181.

Araka, E., Maina, E., Gitonga, R., & Oboko, R. (2020). Research trends in measurement and intervention tools for self-regulated learning for e-learning environments—systematic review (2008–2018). *Research and Practice in Technology Enhanced Learning, 15*(1).

Ceron, J., Baldiris, S., Quintero, J., Garcia, R.R., Saldarriaga, G.L.V., Graf, S., & De La Fuente Valentin, L. (2021). *Self-Regulated Learning in Massive Online Open Courses: A State-of-the-Art Review*. IEEE Access, 9, 511–528.

Devolder, A., Van Braak, J., & Tondeur, J. (2012). Supporting self-regulated learning in computer-based learning environments: systematic review of effects of scaffolding in the domain of science education. *Journal of Computer Assisted Learning, 28*(6), 557–573.

Dong, X., Yuan, H., Xue, H., Li, Y., Jia, L., Chen, J., Shi, Y., & Zhang, X. (2024). Factors influencing college students' self-regulated learning in online learning environment: A systematic review. *Nurse Education Today, 133*, 106071.

Edisherashvili, N., Saks, K., Pedaste, M., & Leijen, Ä. (2022). Supporting Self-Regulated Learning in Distance Learning Contexts at Higher Education Level: Systematic Literature Review. *Front. Psychol., 12*, 792422.

Gambo, Y., & Shakir, M.Z. (2021). Review on self-regulated learning in smart learning environment. *Smart Learning Environments, 8*(1), 12. https://doi.org/10.1186/s40561-021-00157-8

Garcia, R., Falkner, K., & Vivian, R. (2018). Systematic literature review: Self-Regulated Learning strategies using e-learning tools for Computer Science. *Computers & Education, 123*, 150–163.

Goda, Y., Yamada, M., Matsuda, T., Kato, H., Saito, Y., & Miyagawa, H. (2022). From Adaptive Learning Support to Fading Out Support for Effective Self-Regulated Online Learning. *Research Anthology on Remote Teaching and Learning and the Future of Online Education, 254*–274.

Harris, J., Grandgenett, N., & Hofer, M. (2010). *Testing a TPACK-Based Technology Integration Assessment Rubric*. In D. Gibson & B. Dodge (Eds.), Proceedings of Society

for Information Technology & Teacher Education International Conference 2010 (pp. 3833-3840). Chesapeake, VA: AACE

Heikkinen, S., Saqr, M., Malmberg. J., & Tedre, M. (2022). Supporting self-regulated learning with learning analytics interventions – a systematic literature review. *Education and Information Technologies*, *28*, 3059–3088. https://doi.org/10.1007/s10639-022-11281-4

Järvelä, S., Nguyen, A., & Molenaar, I. (2023). Advancing SRL research with artificial intelligence. Computers in Human Behavior, 147, 107847.

Jivet, I., Scheffel, M., Drachsler, H., & Specht, M. (2017). *Awareness is not enough: Pitfalls of learning analytics dashboards in the educational practice*. In European conference on technology enhanced learning (pp. 82-96). Springer, Cham.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130 144.

Lodge, J.M., & Harrison, W.J. (2019). The Role of Attention in Learning in the Digital Age. *The Yale Journal of Biology and Medicine, 92*(1), 21–28.

Matcha, W., Uzir, N. A., Gašević, D., & Pardo. A. (2020). A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Transactions on Learning Technologies, 13*(2), 226-245.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46.

Mirriahi, N., Joksimović, S., Gašević, D., et al. (2018). Effects of instructional conditions and experience on student reflection: A video annotation study. *Higher Education Research and Development, 37*(6), 1245–59

Molenaar, I., Mooij, S.D., Azevedo, R., Bannert, M., Järvelä, S., & Gašević, D. (2023). Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data. *Computers in Human Behavior, 139*, 107540.

Moskal, M., & Leydens, J. (2019). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research, and Evaluation, 7*, 10.

OECD (2019). *Getting skills right: Future-ready adult learning systems*. Paris: OECD Publishing.

Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research. *Front. Psychol*, *8*, 422.

Panadero, E., Klug, J., & Järvelä, S. (2016). Third wave of measurement in the self-regulated learning field: When measurement and intervention come hand in hand. *Scandinavian Journal of Educational Research, 60*(6), 723–735.

Pérez-Álvarez, R., Maldonado-Mahauad, J., & Pérez-Sanagustín, M. (2018). *Tools to Support Self-Regulated Learning in Online Environments: Literature Review*. In Lecture notes in computer science (pp. 16–30).

Pintrich, P.R. (2000). *The role of goal orientation in self-regulated learning*. In M. Boekaerts, P.R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 451–502). Academic Press.

Radović, S. (2024). Is it only about technology? The interplay between educational technology, teaching practice, and students' learning activities. *Journal of Computers in Education, 11*, 743–762.

Radović, S., & Seidel, N. (2024a). Introducing the SRL-S rubric for evaluating technology-enhanced learning environments for self-regulated learning. Accepted for publication in *Innovative Higher Education Journal*.

Radović, S., & Siedel, N. (2024b). *Bridging learning science and learning analytics: Self-Regulation Learning support (SRL-S) rubric*. 14th International Conference on Learning Analytics & Knowledge (LAK24). 18. – 22. 3. 2024, Kyoto, Japan.

Radović, S., Seidel, N., Haake, J.M., & Kasakowskij, R. (2024a). Analyzing students' self-assessment practice in a distance education environment: Student behavior, accuracy, and task-related characteristics. *Journal of Computer Assisted Learning, 40*(2), 654–666.

Radović, S., Seidel, N., Menze, D., & Kasakowskij, R. (2024b). Investigating the effects of different levels of students' regulation support on learning process and outcome: In search of the optimal level of support for self-regulated learning. *Computers & Education*, *215*, 105041.

Reddy, M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448.

Sghir, N., Adadi, A., & Lahmer, M. (2022). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies, 28*(7), 8299–8333.

Thaler, N., Kazemi, E., & Huscher, C. (2009). Developing a rubric to assess student learning outcomes using a class assignment. *Teaching of Psychology*, *36*(2), 113–116.

Viberg, O., Khalil, M., & Baars, M. (2020). *Self-regulated learning and learning analytics in online learning environments*. In Proceedings of the 10th International Conference on Learning Analytics Knowledge (LAK'20). ACM, New York, NY, USA, 11 pages.

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2023). When adaptive learning is effective learning: comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments, 31*(2), 793–803.

Wu, T.T., Lee, H.Y., Li, P.H., Huang, C.N., & Huang, Y.M. (2023). Promoting Self-Regulation Progress and Knowledge Construction in Blended Learning via ChatGPT-Based Learning Aid. *Journal of Educational Computing Research*, *61*(8), 3–31.

Zimmerman, B.J. (2000). *Attaining self-regulation: a social cognitive perspective*. In Handbook of Self-Regulation, eds M. Boekaerts, P.R. Pintrich, and M. Zeidner (San Diego, CA: Academic Press), 13–40. Altman, D.G. (1999). Practical statistics for medical research. New York, NY: Chapman & Hall/CRC Press.

## APPENDIX

**Appendix A.** The SRL-S rubric - assessing the extent of SRL support within a learning environment

| SRL | | Limited SRL support (1) | Moderate SRL support (2) | Advanced SRL support (3) |
|---|---|---|---|---|
| Forethought Phase | *F1. Goal Setting* | LE provides goals predefined by the teacher, and do not allow students to set or modify their goals within LE, nor can they easily access goal related performance indicators. | LE offers detailed insights into students' learning progress in relation to the course goals. However, it does not allow students to set or modify their own learning goals within the platform. | LE offers a variety of learning goals for students to choose from (e.g. course mastery or passing). Students can also set custom goals related to content or performance. Also, LE provides detailed analysis related to the chosen goals. |
| | *F2. Strategic Planning* | LE facilitates the sharing and accessibility of learning resources but does not include tools to help students select learning paths, determine appropriate actions, or plan task execution. | LE provides with an overview of all available learning resources (those completed, left unfinished, or which are next), allowing them to quickly access, prioritize tasks and identify the materials they need. | LE provides students with an overview of all available learning resources, along with useful information such as success rates, progress tracking, and estimated time required for each resource. |
| | *F3. Self-Efficacy and Outcome Expectation* | LE provides minimal information, typically at midterm, about students' past performance, such as their success, progress, effort, or time spent. It does not actively promote the development of self-efficacy. | LE offers detailed information about students' performance, progress, and effort, while also prompting them to reflect on their self-perceived efficacy and assess their capabilities. | LE provides students with details about their efficacy or prompts them to reflect on their self-perceived efficacy. LE goes a step further by offering predictions (about success, outcomes, time needed, etc.) and help to set realistic expectations. |
| | *F4. Task Value and Interest* | LE provides assignments with no or limited practical application, connection to next learning chapters, or other subject or courses. | LE allows students to apply their knowledge to solve realistic practice assignments (follows the principles of authenticity). | LE provides advanced learning technologies that allows students to use professional tools, skills, or relevant methods (for their study or selected goal) to create or self-assess knowledge. |
| | *F5. Goal Orientation* | LE provides only general information regarding the course requirements (goal set by teacher). Students lack visibility into how they are performing or advancing towards their goals. | LE offers students' detailed criteria for success and displays their performance in relation to the goal (set by teacher). Students can compare progress and performance against the criteria and their goal. | LE goes beyond providing information about students' progress, process, and outcome in relation to their goals. It also visualizes what and how needs to be improved or adjusted to attain the selected goal. |
| Performance Phase | *P1. Self-Instruction* | LE provides outline and table of learning content. Besides, there are general instructions about the course requirements to helps individuals take control of their learning. | LE provides task-specific or general self-questions along learning resources to prompt students to achieve desired outcomes. | LE provides adaptive cues that directed cognitive process and thinking during learning. A technology (like intelligent chatbot or similar) uses motivational technique to instruct steps in the coping process. |

| | | | | |
|---|---|---|---|---|
| | ***P2. Imagery*** | LE uses images and visual representation of learning material to support the forming of vivid mental pictures and visual models. | LE includes videos and tools for graphical strategies within the text (annotations, color-coded text, and similar visual aids are utilized to enhance knowledge organization). | LE provides interactive simulations or virtual reality space for developing knowledge and practicing skills. LE could also support students in creating concept maps and visualizations. |
| | ***P3. Time Management*** | LE provides limited support for time management e.g., only mentioning deadlines and exam dates. LE do not record nor analyze time spent on learning. | LE provides information about students' past performance as well as the time spent on specific learning resources and overall learning. Deadlines reminders could be sent. | LE provides information about students' past behavior (or success, progress, time, etc.), but also offers future predictions on managing time effectively in relation to their selected goals. |
| | ***P4. Help Seeking*** | LE facilitates scheduled communication with the teacher. However, it lacks clear avenues or guidance for students to seek assistance when encountering challenges. | LE offers a/synch channels for communication (forum, chat, LMS tools, etc.) which students can use to engage with peers and teachers, to ask questions, share concerns, or request support. | LE instructs and supports students to use various communication channels (e.g., tasks shared with peers, collaborative joint activities). Additionally, help seeking support includes external resources, AI agents, or querying LLM. |
| | ***P5. Task Strategies*** | LE provides a general description of different strategies that can be used. There is no specific structure to support students in performing different tasks. | LE offers task-related support strategies during learning activities (e.g. solving tasks, or self-assessing task solutions). Students are supported in redoing tasks using alternative strategies. | LE offers task-specific strategies for different tasks (this can include tips on critical thinking, summarization, application of skills). Moreover, LE provide feedback on students' learning strategies, behavior, and effective strategies etc. |
| | ***P6. Metacognitive Monitoring*** | LE do not specifically support analytics, monitoring understanding, and evaluating success of chosen learning strategies; aside from providing knowledge tests and tasks that require manually scoring results. | LE supports students in monitoring their progress in relation to general course outcomes. Students can gauge their overall performance (or success, progress, etc.) against the formal objectives of the course (usually via learning dashboards). | LE enables students to compare their progress globally, but also in relation to learning units, specific materials (e.g., texts, tasks, reflections), and individual items. Additionally, LE provides monitoring of SRL behavior, used strategies, and learning patterns. |
| **Self- Reflection Phase** | ***S1. Self-Evaluation*** | LE provides a sample solution that may help students to self-evaluate their solution against master solution (feed-up). However, it does not support identification of areas for improvement. | LE provides different types of tasks that allow students to evaluate their knowledge and skills through, for example, various assessments, self-assessments, or quizzes (feed-back). | LE provides tasks for self-evaluation, but also offers additional data analysis, feedback and guidance on specific areas that need attention (feed-forward). LE marks and visualizes tasks that are underperformed in relation to the goal, recommending improvement. |
| | ***S2. Causal Attribution*** | LE offers a limited resources to reflect (e.g., knowledge tests and related rubrics). No questions specifically guide students how to evaluate factors of failure. | LE encourages students to consider factors that influenced their failures. For example, self-assessment tasks involve rating solutions against different criteria or master solution. | LE includes prompted critical reflection tasks after significant learning events or units. These tasks encourage reflection on strengths and weaknesses, performance, and progress toward achieving goals |

| | | | | |
|---|---|---|---|---|
| | ***S3. Self-Reactions*** | LE includes knowledge assessments with corresponding rubrics, but it does not consider experiences, emotions, or future goals. | LE incorporates learning dashboard that provide insights (awareness and reflection) on their learning activities. | LE provide a learning dashboard together with critical reflection tasks that specifically ask students to reflect on their learning experiences or think about their feelings of satisfaction or disappointment. |
| | ***S4. Adaptation*** | LE provides limited guidance or resources to assist students in modifying or adapting their approaches to learning. This is usually organized as scheduled virtual cohort meetings with teachers. | LE provides information about learning progress and outcome. However, learning material do not adapt, and students cannot directly modify their learning goals within the LE. | LE includes critical reflection tasks that specifically ask students to reflect on adjusting their learning strategies, setting new goal within LE, and to adapt their strategies (based on the information about learning progress ). |

**Note:** *As introduced in Radović and Seidel (2024a). Assign performance levels to each criterion. The corresponding rating (1, 2, or 3) can be assigned only if all requirements from the level are fulfilled. Otherwise, a lower rating should be given (except for when "limited" level has not been reached, then 0 should be given*

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 699–720
https://doi.org/10.21449/ijate.1401339
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Development and validation of STEM motivation scale for middle school students

**Arif Açıksöz**[1*], **İlbilge Dökme**[2], **Emine Önen**[3]

[1]Republic of Türkiye Ministry of National Education, Konya, Türkiye
[2]Gazi University, Gazi Faculty of Education, Department of Science Education, Ankara, Türkiye
[3]Gazi University, Gazi Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

**ARTICLE HISTORY**

*Received: Dec. 06, 2023*

*Accepted: Aug. 26, 2024*

**Keywords:**
Expectancy value theory,
Motivation,
Scale development,
STEM.

**Abstract:** Understanding motivational beliefs such as expectancy and value that shape students' persistence and decision to pursue a STEM career, obtaining valid and reliable measures for these dimensions, and developing strategies using this data are critically important to ensure students' persistence in the STEM pipeline. Therefore, this study aims to develop a tool to measure middle school students' STEM motivations within the expectancy and value concepts framework. The trial version of the scale was conducted on 967 middle school students in the 5th, 6th, 7th, and 8th grades. The study group was randomly divided into two groups. EFA was conducted on the data obtained from the first sub-group (n=479), and CFA was performed using the data obtained from the second sub-group (n=488). The results of a series of CFA performed to test three different models developed based on the theoretical structure, Model 3, the second-order single-factor structure composed of 5 sub-dimensions was found to be a successful model. This measurement tool would allow determining motivational beliefs within the expectancy-value concept that can be targeted to encourage students' interest in STEM fields, as well as help design interventions for these structure(s), and evaluate the effectiveness of these interventions.

## 1. INTRODUCTION

In the last few decades, as technological and industrial advances have accelerated, the demand for STEM (Science, Technology, Engineering, and Mathematics) workforce has begun to increase markedly. Since the number of jobs that require STEM knowledge and skills is rising (Langdon et al., 2011), more STEM professionals are needed to meet this increasing demand (Ball et al., 2017; Hermans et al., 2022; Razali, 2021). Accordingly, STEM education, which refers to teaching and learning in the fields of science, technology, engineering, and mathematics (Gonzalez & Kuenzi, 2012), is considered an important approach to meeting STEM workforce demands for the competitive world of the 21st century (Breiner et al., 2012; Çorlu et al., 2014; Kuenzi, 2008; Kuo et al., 2019; Luo et al., 2021; National, Research Council [NRC], 2011; National Science and Technology Council [NSTC], 2018; PCAST, 2010).

Despite STEM education being widely recognized as crucial for societal advancement and human development, recent reports indicate a decline in the number of students pursuing STEM

*CONTACT: Arif AÇIKSÖZ ✉ arifaciksoz@gmail.com Republic of Türkiye Ministry of National Education, Konya, Türkiye



**e-ISSN: 2148-7456**

majors and entering STEM careers (pipeline problem in STEM) (Griffith, 2010; Hinton Jr. et al., 2020; Sanders, 2009; van den Hurk et al., 2019; Yahaya et al., 2022). Too many students lose their interest in science and mathematics at early ages and make an early exit from the STEM pipeline (Sanders, 2009). Students' reluctance to pursue a STEM career or decline in interest in STEM careers are considered a major STEM problem in many parts of the world (Bøe et al., 2011; Hossain & Robinson, 2012; Perez et al., 2019).

Although studies examining possible reasons for the decline in STEM interest in the last decade highlighted many factors (see, van den Hurk et al., 2019; Wang & Degol, 2013), psychological studies have revealed that it is partly an issue of motivation (Rozek et al., 2017). Motivation refers to the power that stimulates an organism to start and act toward a specific behavior, and explains the intensity, direction, and persistence of this behavior (Petri & Govern, 2012). In the previous studies, some motivation-related factors such as interest, perceived value, feeling competent in STEM disciplines, belief in success, and considering STEM topics as personally interesting and important were found to affect students' willingness to pursue a STEM career (Perez et al., 2019; Robnett & Leaper, 2012). Students' motivation for STEM can be therefore argued to play an important role in interest and continuous engagement in this field, as well as in choosing a STEM career (Chen & Dede, 2011; Joseph et al., 2019; Luo et al., 2019, Robnett & Leaper, 2012; Wang, 2013; Wang & Degol, 2013).

Motivation researchers have introduced many theories based on internal and external factors to explain how motivation affects one's choices, determination, and performance (Bandura's Self-efficacy theory, Covingtoh's Self-worth theory, Ryan and Deci's Self-determination theory, Weiner's Attribution theory, Eccles-Parsons et al.'s Expectancy-value theory, etc.) Among these contemporary educational psychology theories, the Expectancy-value theory (EVT) is particularly focused on the relation of beliefs, values, and goals to actions (Eccles & Wigfield, 2002). Therefore, the EVT has inspired many education-related studies and practices for more than one-quarter of a century (Trautwein et al., 2012).

### 1.1. Expectancy – Value Theory (EVT)

EVT is an important theory developed to understand individuals' motivational beliefs (Eccles & Wigfield, 2020), which is widely used in education to explain and predict students' achievement, persistence, and aspirations (Loh, 2019). This theory assumes that students' motivation to perform achievement tasks (e.g., an effort to do homework or exhibit a skill, engaging in specific activities, or using strategies to develop skills) is determined by their expectation of success in a task and the value they attached to the task (Dotterer, 2022; Rosenzweig et al., 2019). In simpler terms, individuals' motivation for success is a function of their belief in their abilities and the value they place on the task (Wigfield et al., 2009).

Among the components of the EVT, the expectation of success is defined as individuals' beliefs about how well they will perform in future achievement tasks (Meyer et al., 2019; Rosenzweig et al., 2019; Wigfield & Gladstone, 2019). In this context, one's expectations for success predict achievement-related factors including performance, persistence, and choices. For example, when students believe that they are competent in mathematics and expect their successes to continue, they are likely to show good performance in mathematics (Eccles et al., 1983). On the other hand, students with low expectations are more likely to procrastinate on academic tasks (Wu & Fan, 2017).

According to EVT, an individual's expectations of success in any task are strongly influenced by his/her confidence in performance (self-efficacy) or beliefs about his/her ability to perform the task (self-concept beliefs) (French et al., 2023). Ability beliefs are children's evaluations of their current competencies or abilities (Wigfield & Gladstone, 2019). Therefore, many researchers in the field of EVT combine beliefs regarding skills with expectancy values rather than simply measuring expectations (Rosenzweig et al., 2019). Although they have different origins, many empirical studies have also shown that expectations overlap with self-

efficacy (Appianing & Van Eck, 2018). Self-efficacy refers to one's beliefs about their performance in events that affect their life. These beliefs that they can complete a particular task are important predictors of activity choices, willingness to expend effort, and persistence (Bandura, 1997). Thus, scholars sometimes measure self-efficacy instead of expectations or beliefs about skills (Wigfield & Eccles, 2000).

Another component of the EVT, subjective task value, refers to the quality of a task or activity that increases or decreases the probability of being selected by the person (Eccles & Harold, 1991). The incentives during the performance of the task are associated with this component (Gråstén, 2016). When a task is perceived as motivating (seen as important, beneficial, enjoyable, etc.) from an individual's perspective, the likelihood of that task being completed increases (Barron & Hulleman, 2015; Schoenherr, 2024). Conversely, when there is no reason or incentive for the task, it leads to the task not being done (Wentzel & Wigfield, 1998).

Task values vary depending on task characteristics and their impact on the individual's motivation to complete the task. The values, therefore, are unique to the task (Eccles et al., 1983; Wigfield & Eccles, 1992). These values are also subjective because beliefs about an activity are students' own beliefs, and every student is different (Wigfield & Cambria, 2010). For example, success in mathematics is valuable for some students, whereas it might not be valuable for other students (Eccles, 2011). Subjective task value is positively affected by three components namely, attainment value (importance), intrinsic value (interest), and utility value, whereas it is negatively affected by cost value (Eccles et al., 1983; Eccles, 2005; Rosenzweig et al., 2019; Wigfield et al., 2017; Wigfield & Gladstone, 2019).

Eccles et al., (1983) defined attainment value (importance) as the personal importance attributed to succeeding in a task. For example, learning to play a new instrument can be a way for a musician to improve his/her musical skills. In this case, the attainment value of learning to play a new instrument will be high for the musician. In addition, this value is related to one's self-identity (Eccles, 2005). Tasks are considered important when they are consistent with one's self-scheme, gender, ethnicity, and other personality traits or when the task allows one to express their important aspects or affirm themselves (Eccles, 2011; Wigfield et al., 2009; Wigfield & Eccles, 2002). If one wants to affirm him/herself with a task that requires skills or effort, the attainment value of this task increases (Eccles & Harold, 1991).

Intrinsic value refers to the natural and immediate pleasure experienced by an individual during engagement in an activity or their subjective interest in that activity (Eccles & Wigfield, 1995; Partridge, 2013; Wigfield et al., 2009). For example, if a student shows interest in activities carried out in a lesson and finds them entertaining, this student's intrinsic value probably increases, and s/he would show more effort in the lesson than other students (Ball et al., 2017; Barutcu, 2017; Yurt, 2016). EVT argues that if the intrinsic value of a task is high, the person will be intrinsically motivated to fulfill this task (Eccles & Wigfield, 2002). In some aspects, this component is similar to intrinsic motivation and interest concepts (Wigfield, 1994). However, it should be considered that these structures are based on different theoretical traditions (Wigfield & Cambria, 2010).

Utility value refers to the perceived benefit of the activity (Wentzel & Wigfield, 1998). In other words, it defines how a task fits one’s future plans (e.g., career goals) (Wigfield, 1994). If one finds the task important for their future goals or receives promotions if it is accomplished, they may engage in it (Shin et al., 2019). For example, an additional foreign language course taken by a student may help enhance their language skills, be more effective in international relations, and expand job opportunities. Therefore, taking an extra foreign language course would be highly beneficial for their future career, resulting in high value of benefit. In a sense, this component includes more “external” reasons such as achieving the desired result (Eccles et al., 1999; Wigfield & Eccles, 2002).

The fourth value proposed in the EVT, the cost value, negatively affects student motivation (Barron & Hulleman, 2015; Meyer et al., 2019). This value is conceptualized in terms of fear of social consequences of the task (such as negative reactions from peers, parents, and colleagues) (Eccles, 2011), fear of failure, concerns about performance, amount of effort required for success, and opportunities lost as a result of a choice (Wigfield & Eccles, 2002). The high cost of a task compared to its benefit may cause the individual to avoid that task (Loh, 2019). For instance, completing a math assignment can be cited as an example of task cost. The student must invest time and energy to complete the assignment, potentially sacrificing other activities. According to EVT, there are three different types of cost: the effort required to succeed in the task, lost time that can be spent on other activities, and negative psychological outcomes related to struggle or failure on the task (Barron & Hulleman, 2015; Eccles et al., 1983).

### 1.2. The Current Study

Due to the growing need to pursue a STEM career, raising a continuous interest in STEM is important (Romine & Sadler, 2016). Previous reports indicated that motivation -an important factor that should be targeted to promote learning- (Williams & Williams, 2011) plays a critical role in educational outcomes (Walters et al., 2016). High motivation not only helps students in the learning process but also leads them to value what they learn and develop an interest in future careers (Beerenwinkel & von Arx, 2016). Accordingly, students' motivations can be targeted to increase their interest in STEM fields (Rosenzweig & Wigfield, 2016).

The middle school period is an important stage for the development of students while getting prepared for a rapidly changing future. Many researchers highlighted the importance of the secondary education stage for improving interest in STEM and choosing a STEM field (Christensen & Knezek 2017; English, 2017; Moreno et al., 2016). The STEM skills acquired in this period paved the way for a successful STEM career (Knezek et al., 2013). Brown et al. (2016) observed that middle school students' STEM beliefs and attitudes changed after experiencing the STEM curriculum. Sadler et al., (2012) found that students' career preferences before starting high school are the most powerful predictor of their career preferences when graduating from high school. Tai et al., (2006) reported that middle school students who are interested in a science career are more likely to graduate with a science degree. Consistent with this, Dabney et al., (2012) found that the probability of choosing a STEM career for a student who is not interested in STEM is significantly lower compared to a student who is interested in STEM since middle school. In this regard, measuring middle school students' motivational beliefs such as expectancy and value which shape their decisions to continue a STEM career, obtaining valid and reliable measurements of these dimensions, and designing interventions based on the obtained data are very important to ensure students' persistence in the STEM pipeline.

Considering the long history of Eccles's EVT which is used to understand students' motivational beliefs, many measurement tools are developed based on this theory for different academic levels (primary school, middle school, high school, college, etc.) and fields (mathematics, English, STEM, physical education, critical thinking, Master's degree, etc.) to measure students' motivations (see Appianing & Van Eck, 2018; Barron & Hulleman, 2015; Eccles & Wigfield, 1995; Valenzuela et al., 2011; Wigfield & Eccles, 2000; Xiang et al., 2003). Scales developed by Eccles et al. from these measurement tools are highly preferred due to their factor structure, good psychometric properties, and ability to show the relationships between success and choice (Wigfield et al., 2009). On the other hand, there are measurement tools -although not based on EVT- using some motivational constructs including expectancy/value structures developed to measure students' motivations (Glynn et al., 2011; Jones, 2009, 2018). After a literature survey, detailed information was obtained on some measurement tools, as shown in Table 1.

**Table 1.** *Information on measurement tools.*

| Developed by | Measurement tool | Sample | Theory | Number of items |
|---|---|---|---|---|
| Eccles & Wigfield (1995) | Children's self and task perceptions in the domain of mathematics | Middle & high school students | Expectancy-Value Theory (Eccles et al., 1983) | 19 |
| Glynn et al. (2011) | Science Motivation Questionnaire II | College students | Bandura's social cognitive theory (Bandura 1977-1986) | 25 |
| Jones (2012/2022) | MUSIC Inventory (Middle/High School Student version) | Middle school students | The MUSIC Model of Academic Motivation (Jones, 2009,2018) | 18 |
| Kosovich et al. (2015) | Expectancy-Value-Cost Scale | Middle school students | Expectancy-Value Theory (Eccles et al., 1983) | 10 |
| Appianing & Van Eck (2018) | Value-Expectancy STEM Assessment Scale (VESAS) | College students | Expectancy-Value Theory (Eccles et al., 1983) | 15 |
| Luo et al. (2019) | STEM Continuing Motivation (STEM-CM) | Middle school students | Continuing motivation Maehr (1976) | 25 |
| Kızılay et al. (2019) | Motivation Scale for STEM Fields | High school students | ARSC model Keller (1979) | 22 |
| Gök (2021) | STEM Attitude and Motivation Survey | Middle school students | --- | 34 |

As seen in Table 1, some tools are developed based on different theories to measure students' motivations at different academic levels. The measurement tool developed by Eccles and Wigfield (1995) measures middle and high-school students' motivations in mathematics. On the other hand, the "Expectancy-Value-Cost Scale" developed by Kosovich et al. (2015) employs expectancy/value and can be adapted for certain content fields such as mathematics and science. Another measurement instrument -although not based on the expectancy/value theory- was developed by Glynn et al. (2011). Science motivation questionnaire-II (SMQ-II) consists of different motivational structures (intrinsic motivation, self-determination, self-efficacy, career motivation, and grade motivation) and is frequently used to measure student motivation in science disciplines (biology, physics, and chemistry). Besides, the music model developed by Jones (2009, 2018) combines different motivation theories -also includes the EVT- and focuses on motivation in a specific event and explains factors motivating one to participate in a specific event in a specific discipline (mathematics, science, etc.). In general, each tool used by researchers to measure student motivation is developed to assess a specific area. Although mathematics or science is a part of STEM, as indicated in many definitions (see Bybee, 2010; Gonzalez & Kuanzi, 2012) STEM is a holistic approach and is composed of the disciplines in its content. Therefore, measurement tools developed for a specific discipline may yield indirect outcomes while measuring motivation in STEM. This is why we focused on STEM motivation for the measurement tool we developed. Additionally, as previously mentioned, the middle school years are a critical period for the development of students' motivational beliefs. It is seen that 5 of the measurement tools given in Table 1 are designed for middle school students. Plus, three of these measurement tools (Eccles & Wigfield, 1995; Jones, 2009, 2018; Kosovich et al., 2015) are designed for a specific discipline (e.g., science, mathematics). Luo et al. (2019) and Gök (2021) developed measurement tools focusing directly STEM motivation of middle school students. However, neither measurement tool was based on the EVT. In this study, unlike the previously mentioned measurement tools, we focus specifically on STEM and use the EVT to assess middle school students' expectancies and

values related to their STEM motivation. We believe that such a tool will make valuable contributions to the existing literature in this field.

This study outlines the development process of a tool based on the concepts of expectancy and value to measure middle school students' STEM motivation. This tool can be used to assess students' STEM motivation, design intervention strategies to retain students in this field, and evaluate the effectiveness of these interventions.

## 2. METHOD

### 2.1. Study Group

The current study involved students who were attending a state middle school in the 2020-2021 academic year in Turkey. The study group consisted of 967 students (316 5th graders; 110 6th graders; 266 7th graders; and 275 8th graders) who voluntarily completed the Turkish version of the trial survey. The study group was randomly divided into two groups for analysis. Exploratory Factor Analysis (EFA) was performed on the data obtained from the first sub-group (n=479) and Confirmatory Factor Analysis (CFA) was conducted on the data collected from the second sub-group (n=488). The gender and grade information of the students in the 1st and 2nd groups are shown in Table 2.

**Table 2.** *Gender and grade information of the students in the 1st and 2nd sub-groups.*

| First sub-group | | | | Second sub-group | | | |
|---|---|---|---|---|---|---|---|
| Grade Level | Gender | *f* | % | Grade Level | Gender | *f* | % |
| 5th Grade | Male | 75 | 15.7 | 5th Grade | Male | 86 | 17.7 |
| | Female | 79 | 16.5 | | Female | 76 | 15.6 |
| 6th Grade | Male | 19 | 3.9 | 6th Grade | Male | 23 | 4.7 |
| | Female | 38 | 7.9 | | Female | 30 | 6.1 |
| 7th Grade | Male | 60 | 12.5 | 7th Grade | Male | 60 | 12.2 |
| | Female | 63 | 13.2 | | Female | 83 | 17 |
| 8th Grade | Male | 70 | 14.6 | 8th Grade | Male | 55 | 11.3 |
| | Female | 75 | 15.7 | | Female | 75 | 15.4 |
| Total | | 479 | 100 | Total | | 488 | 100 |

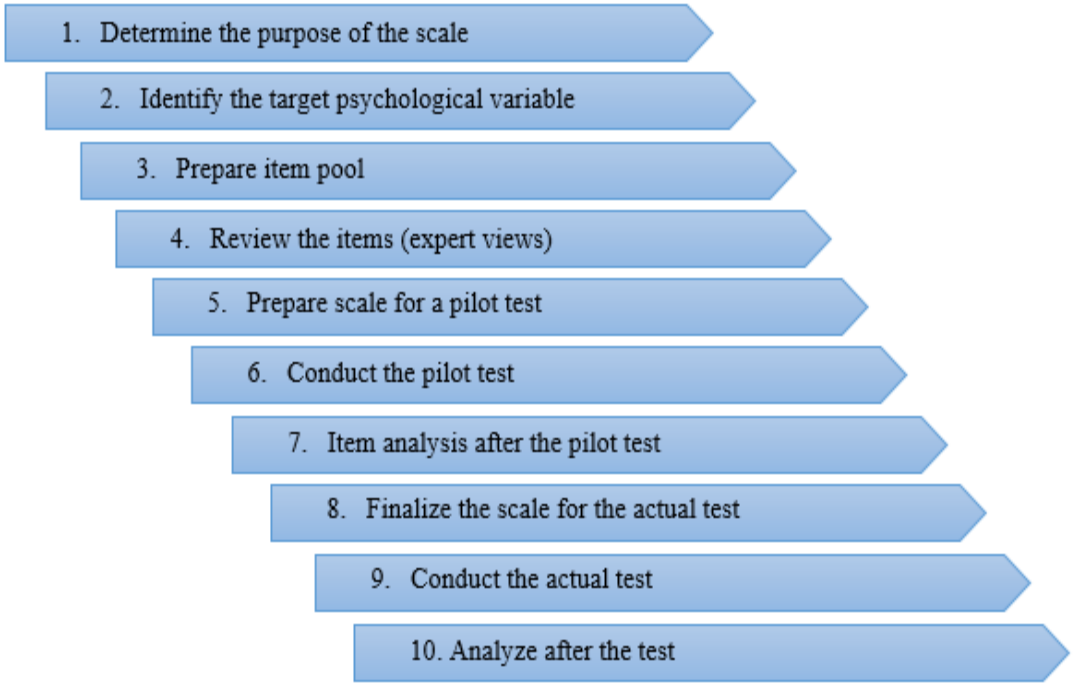
As seen in Table 2, the first sub-group has a balanced distribution of gender in all grade levels (5, 6, 7, 8). The highest-class size in this sub-group was observed in 5th grade with 75 boys and 79 girls (n=154), whereas the lowest class size was in 6th grade with 19 boys and 38 girls (n=57). Similar to the first sub-group, the second sub-group also had a balanced distribution of gender in all grade levels. Plus, as in the first sub-group, the highest-class size was observed in 5th grade with 86 boys and 76 girls (n=162), and the lowest class size was in 6th grade with 23 boys and 30 girls (n=53). In general, both sub-groups had a balanced distribution of gender and grade level.

### 2.2. Scale Development

As shown in Figure 1, the scale development steps proposed by DeVellis (2003, s.60-137) were followed during the scale development study. As mentioned before, it is an interesting fact that STEM motivation is an important factor to retaining students in a STEM field, and accordingly, measuring directly STEM motivations instead of motivation in each discipline (mathematics, science, etc.) is considered important by the researchers. Therefore, this study was aimed at developing a measurement tool for secondary students' STEM motivations. Accordingly, to measure middle school students' STEM motivations, the EVT introduced by Eccles et al. (1983) was studied in detail, and comprehensive definitions of the components of this theory were made. Then considering the scales developed based on the expectancy-value theory and components of the theory of Eccles et al. (1983), 35 items were prepared with a 5-point Likert-

type scale (1: Strongly Disagree, 2: Disagree, 3: Neutral, 4: Agree, 5: Strongly Agree). Some examples of the scale items are presented in Table 3.

**Figure 1.** *DeVellis's scale development steps*




Expert opinions were received to determine whether the items were appropriate to measure the intended characteristic. Accordingly, the items were sent to 2 assessment and evaluation experts, 2 STEM experts, 1 expert studying STEM motivation, and 3 doctoral students who received STEM education for review. Two assessment and evaluation experts, 2 STEM experts, and 3 doctoral students provided opinions on the items. Based on the expert opinions, 2 items were removed from the scale, and 2 items were revised. Then the revised version of the scale consisting of 33 items (six cost items were negatively worded) was examined by a language expert and then by two science teachers regarding language and understandability. The scale was then decided to correct in spelling-grammar and is understandable for middle students. However, to further test the understandability, the scale was applied to a small group of 5th graders (n=30). Before the application, the students were informed about the scale and it was stated that the definition of STEM field, STEM field professions, and courses in the scale were explained at the bottom of the scale. As the students did not have any understandability issues while responding, the scale was decided to be understandable and ready for implementation.

**Table 3.** *Some examples for the scale items.*

| Dimension | Item No. | Item with English |
|---|---|---|
| Expectancy | Item 1 | STEM alanlarında diğer alanlara kıyasla daha başarılı olacağıma inanıyorum.<br>[I believe I will be more successful in STEM fields than in other disciplines.] |
| Attainment value | Item 12 | STEM alanlarında öğreneceklerimi önemsiyorum.<br>[I care about the things I learn in STEM fields.] |
| Utility value | Item 16 | STEM alanlarına yönelik öğrendiklerim iyi bir meslek sahibi olmamı sağlayacaktır.<br>[Things I learn in STEM fields will allow me to gain a good profession.] |
| Intrinsic value | Item 21 | STEM ile ilgili etkinlikler eğlencelidir.<br>[STEM-related activities are fun.] |
| Cost value | Item 31 | STEM ile ilgili bir etkinliğe zamanımı harcamak istemem.<br>[I don't want to spend my time in a STEM-related activity.] |

After receiving the required ethical permission to conduct the study, the scale was applied to students who voluntarily agreed to participate in the study. Following this practice, the study group composed of volunteer students was randomly divided into two subgroups. EFA was conducted for the pilot study using the data obtained from the first sub-group. On the other hand, CFA was performed for the actual study using the data obtained from the second sub-group.

### 2.3. Data Analysis

To examine the psychometric properties of motivation measures obtained from the developed scale, analysis was conducted using IBM SPSS Statistics version 22.0 and LISREL version 8.8. Before the analysis, the negatively worded items (cost items) were reversely scored. Furthermore, missing data were examined by Little's MCAR (Missing Completely at Random) test. The results of the test conducted on the dataset showed the dataset contains random patterns ($\chi^2$=1955.839, $p$<.000) (Garson, 2015). Accordingly, it was decided that the missing data would not lead to problems in analysis, and assignments were made using the EM algorithm for missing data. Afterward, the study group was randomly divided into two sub-groups to examine the psychometric properties of the scale. To get evidence related to the construct validity of the measures, EFA was conducted on the data obtained from the first sub-group using direct oblimin rotation (since the structures of the theory are correlated) with SPSS ver. 22.0. Since it is the commonly used method in Social Science, Principal Component was used as the factor-extracting method in this study. The appropriateness of these data for EFA was assessed based on the Kaiser criterion (Kaiser, 1960) and Bartlett's Test of Sphericity. Additionally, scree plots and interpretability criteria were used to determine the number of factors.

The EFA conducted for the factor structure of the scale and the second-order factor model developed based on the EVT from the 33-item scale were evaluated together. According to the results of these evaluations (discussed in the next section), the scale was revised to 27 items. Alternative first-order and second-order measurement models were defined based on the factor structure of the 27-item scale and were tested by a series of CFA using the data obtained from the second sub-group. In the model specification, for each latent variable, one-factor loading per latent variable was fixed to 1. Before CFA, to test the multivariate normality assumption, z values for Multivariate Kurtosis (z=26.723, $p$<.000) and skewness (z=57.258, $p$<.000) were calculated. $\chi^2$ value ($\chi^2$=3992.596, $p$<.000) for Multivariate Kurtosis and skewness was also computed. The results indicated that the dataset does not meet the multivariate normality assumption. Accordingly, for parameter estimation, the Robust Maximum Likelihood method was used. Accordingly, the Satorra-Bentler $\chi^2$(S-B$\chi^2$) value was calculated and evaluated (Brown, 2006, s.76). In the CFA, an adequate fit of the measurement models to the data (GFI≥.90, CFI≥.95, NFI≥.90 & RMSEA≤.08) was assessed as evidence for construct validity (Schermelleh-Engel et al., 2003). Both for EFA and CFA, items with loadings higher than .32 were considered an appropriate indicator of the measured construct (Tabachnick & Fidell, 2007). On the other hand, in EFA, items loaded on two or more factors with loadings greater than .10 were considered cross-loading. As evidence for the reliability of these measures, Cronbach's alpha values were calculated using SPSS software version 22.

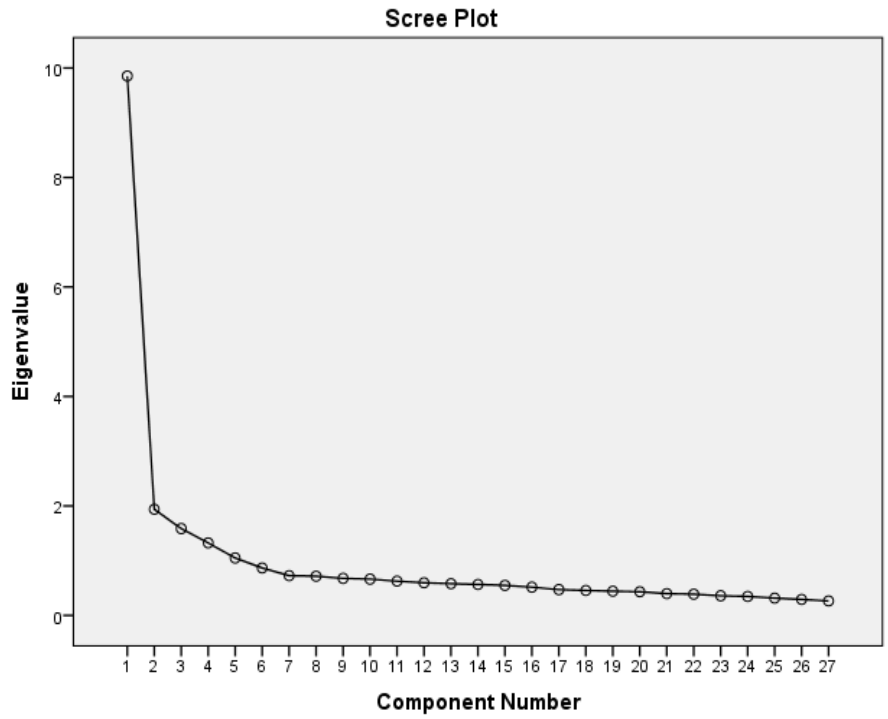
## 3. RESULTS

Firstly, EFA was conducted on the data obtained from the first sub-group (n=479). The KMO value (KMO=.949) and results of Bartlett's test of sphericity ($\chi^2$=6831.4, $p$≤.05) indicated that EFA is feasible for this dataset. The EFA results supported a 5-factor solution, and these 5 factors explained 53.945% of the total variance. However, for one item, the main loading was found to be below .32, and five items had cross-loadings. Therefore, the 9th item was removed since it had the lowest factor loading (λ=.29), and EFA was conducted again. The analysis results showed that items 13, 28, and 29 did not load their expected factor, they rather loaded another factor with a higher loading value. These items were, therefore, removed from the scale,

each time one item, and another EFA was performed after the removal of each item. Then items 14 and 18 were removed, respectively since these cross-loaded factors cause a high inter-correlation between factors, prevent the discrimination of factors, and make it difficult to determine the factor structure. An EFA was conducted again after the removal of each item.

After item removal procedures, a final EFA was conducted on the 27-item scale (KMO=.946, for Bartlett's Test of Sphericity $\chi^2$=5653.59, $p \leq .05$), and a 5-factor solution with an eigenvalue greater than 1 was obtained (Figure 2).

**Figure 2.** *Scree-plot graph for a 5-factor solution*




The eigenvalue of the first factor is 9.854, and it explains 36.496% of the total variance. However, the eigenvalues of the remaining 4 factors varied between 1.049-1.939, and each of these factors explains only a small amount of variance. The eigenvalues, explained variance, and factor loadings are shown in Table 4.

Considering the theory and the items loaded on the factors, the first factor was called “Intrinsic value", the second factor was "Cost", the third factor was "Utility", the fourth factor was "Expectancy", and the fifth factor was called as "Attainment”. All factor loadings were above .32. The factor loadings values varied between λ=.627 and .844 for the first factor; between λ= .530 and .769 for the second factor; between λ=.422 and .753 for the third factor; between λ= -.480 and -.798 for the fourth factor; and finally, varied between λ=-.529 and -.638 for the fifth factor. The EFA results indicate that the final version of the scale consisting of 27 items can measure middle school students’ STEM motivation over the expectancy, utility value, attainment value, intrinsic value, and cost dimensions defined in the theory.

In addition to the EFA, the validity of the measurements obtained from the scale was tested with CFA conducted on the expectancy-value model (Model 1). This model was developed based on the EVT. As explained in the EVT section, expectancy for success and task value are the two main components of this theory. On the other hand, according to the EVT, task value is positively affected by three factors namely, attainment/importance value, intrinsic value, and utility value (usefulness of the task), whereas, is negatively affected by cost value (Eccles, 2005; Rosenzweig et al., 2019; Wigfield et al., 2017). Accordingly, in the second-order factor model, expectancy and value were higher-order factors; Intrinsic value, Cost, Utility, and

Attainment were first-order factors, and the related items were defined as indicators. According to the calculated fit indexes ($\chi^2$=850.90, *df*=491, GFI=.88, NFI=.96, CFI=.98, and RMSEA=.039), the model showed a good fit to the data. However, the examination of individual parameter estimates (standardized solution) showed that higher constructs were highly correlated, and the Heywood case was observed for the coefficient (β=1.01) indicating the predictive strength of the value higher construct for the attainment first-order construct.

**Table 4.** *EFA analysis results.*

| Factors | Items | F1 | F2 | F3 | F4 | F5 | Eigenvalue | Explained variance |
|---|---|---|---|---|---|---|---|---|
| Intrinsic value | 26 | **.844** | .030 | -.018 | .032 | .065 | 9.854 | 36.496% |
| | 25 | **.762** | -.029 | -.015 | -.055 | -.030 | | |
| | 22 | **.761** | .113 | -.055 | -.049 | -.103 | | |
| | 21 | **.712** | .053 | -.052 | -.062 | -.042 | | |
| | 24 | **.637** | -.005 | .097 | .006 | -.071 | | |
| | 23 | **.636** | -.106 | .037 | -.061 | -.201 | | |
| | 27 | **.627** | -.021 | .089 | -.143 | .183 | | |
| Cost value | 32 | .021 | **.769** | -.015 | -.099 | -.069 | 1.939 | 7.128% |
| | 30 | -.095 | **.759** | .002 | -.016 | .136 | | |
| | 33 | .033 | **.730** | -.073 | -.211 | .023 | | |
| | 31 | .361 | **.530** | .116 | .207 | -.189 | | |
| Utility value | 19 | -.051 | -.071 | **.753** | -.145 | .046 | 1.586 | 5.873% |
| | 17 | -.008 | .036 | **.711** | .018 | .022 | | |
| | 20 | .046 | .046 | **.698** | .012 | -.135 | | |
| | 16 | .033 | -.010 | **.639** | -.044 | -.196 | | |
| | 15 | .313 | -.062 | **.422** | -.132 | .138 | | |
| Expectancy | 3 | -.032 | .035 | .065 | **-.798** | .038 | 1.325 | 4.906% |
| | 1 | .010 | .020 | .052 | **-.768** | -.047 | | |
| | 7 | .211 | -.081 | .091 | **-.655** | .111 | | |
| | 2 | .116 | .092 | -.066 | **-.643** | -.183 | | |
| | 4 | .140 | .083 | .006 | **-.588** | -.071 | | |
| | 6 | -.023 | .153 | .142 | **-.565** | -.157 | | |
| | 5 | .131 | .231 | .067 | **-.480** | -.088 | | |
| Attainment value | 12 | .112 | .071 | .210 | .013 | **-.638** | 1.049 | 3.885% |
| | 11 | .057 | .094 | .286 | -.051 | **-.613** | | |
| | 10 | .064 | -.159 | -.169 | -.263 | **-.572** | | |
| | 8 | .037 | .021 | .324 | -.110 | **-.529** | | |

After the 33-item version of the scale was determined to be not successful by the CFA, a revised scale consisting of 27 items was obtained using EFA results. In addition to Model 1 defined for analysis of the 33-item scale, two measurement models (Model 2 and Model 3) were also defined and CFA analyses were conducted on these models using the data obtained from the 2nd sub-group. Accordingly, a 5-factor measurement model, Model 2 (expectancy, intrinsic value, utility value, attainment value, and cost value were considered factors, and the items were considered indicators) consistent with the 5-factor solution obtained by EFA was defined and tested. However, the EFA results indicated that the variance explained by the intrinsic value factor was 36.496%, and there might be other structure(s) over the determined factors. The sub-dimensions (utility, attainment, cost, and intrinsic values) under the expectancy and value constructs of the theory are often highly correlated with each other or loaded on a factor (Eccles & Wigfield, 1995). Furthermore, Trautwein et al. (2012) found strong relations between

expectancy and value beliefs. It is, therefore, highly possible that strong relations exist between expectancy and value as well as between the sub-dimensions of value. Accordingly, another second-order factor model (Model 3) based on the EVT was defined and tested. In this model, expectancy, intrinsic value, utility value, attainment value, and cost value were considered first-order factors; motivation was a second-order factor, and the items were considered indicators. Fit statistics for the models developed based on the 27-item scale are shown in Table 5.

**Table 5.** *Model fit indices for the tested models (27-item scale).*

| Model | Chi-Square | *df* | GFI | NFI | CFI | RMSEA |
|---|---|---|---|---|---|---|
| Model 1 | 586.64 | 320 | .89 | .96 | .98 | .041 |
| Model 2 | 477.69 | 314 | .91 | .97 | .99 | .033 |
| Model 3 | 515.87 | 319 | .91 | .96 | .99 | .036 |

As seen in Table 5, the fit indices of Model 1 (obtained based on the 27-item scale) display an acceptable fit to the data. According to the test of the model, factor loading estimates (λ's) and unique variances (ε's) vary between .49-.85 and .28-.76, respectively. On the other hand, the examination of the correlations between latent variables (see Table 6) indicated a strong correlation between expectancy and value factors ($r$=.94; $p$<.05). Furthermore, the evaluation of individual parameter estimates (standardized solution) showed that higher-constructs were highly correlated with each other, and Heywood case was observed for the coefficient (β=1.03) indicating the predictive strength of value higher-construct for the attainment first-order construct. Accordingly, Model 1 was decided to be not consistent with the measures obtained from the 27-item scale.

**Table 6.** *Correlation matrix for Model 1.*

| | Exp. | Value | Att. | Uti. | Int. | Cost |
|---|---|---|---|---|---|---|
| Exp. | 1.00 | | | | | |
| Value | .94 | 1.00 | | | | |
| Att. | -- | .98 | 1.00 | | | |
| Uti. | -- | .86 | .85 | 1.00 | | |
| Int. | -- | .86 | .85 | .74 | 1.00 | |
| Cost | -- | .70 | .69 | .60 | .60 | 1.00 |

Similar to Model 1, Model 2 (obtained based on the 27-item scale) also showed an acceptable fit to the data. The factor loading estimates (λ=.42 - .75; $p$<.05) obtained by the test of the model pointed out that the indicators of this model are accurate indicators of the constructs and dimensions of the model. However, the correlations between latent variables (see Table 7) varied between .37-.83. Furthermore, high correlations were found between expectancy value and attainment value ($r$=.83; $p$<.05); and between attainment value and utility value ($r$=.83; $p$<.05). These findings indicated that the dimensions of the scale do not discriminate well, and the model do not represent the factor structure of the measures sufficiently.

**Table 7.** *Correlation matrix for Model 2.*

| | Exp. | Att. | Uti. | Int. | Cost |
|---|---|---|---|---|---|
| Exp. | 1.00 | | | | |
| Att. | .83 | 1.00 | | | |
| Uti. | .66 | .83 | 1.00 | | |
| Int. | .74 | .66 | .57 | 1.00 | |
| Cost | .54 | .47 | .40 | .59 | 1.00 |

Like other models, Model 3 (obtained based on the 27-item scale) also displayed an acceptable fit to the data. Standardized estimates for both factor loadings (λ's = .42–.75) and unique variances (ε's = .44–.83) indicated that the items of the 27-item scale are appropriate indicators of their respective factors and can produce measures with acceptable levels of error. On the other hand, in addition to the evidence of construct validity for the measures, the coefficients indicating the predictive strength of the latent variable in the model (see Table 8) for 5 factors were found to be high (they varied between .60-.91). Second-order measurement model (Model 3) with standardized solutions is shown in Figure 3.

**Table 8.** *Correlation matrix for Model 3.*

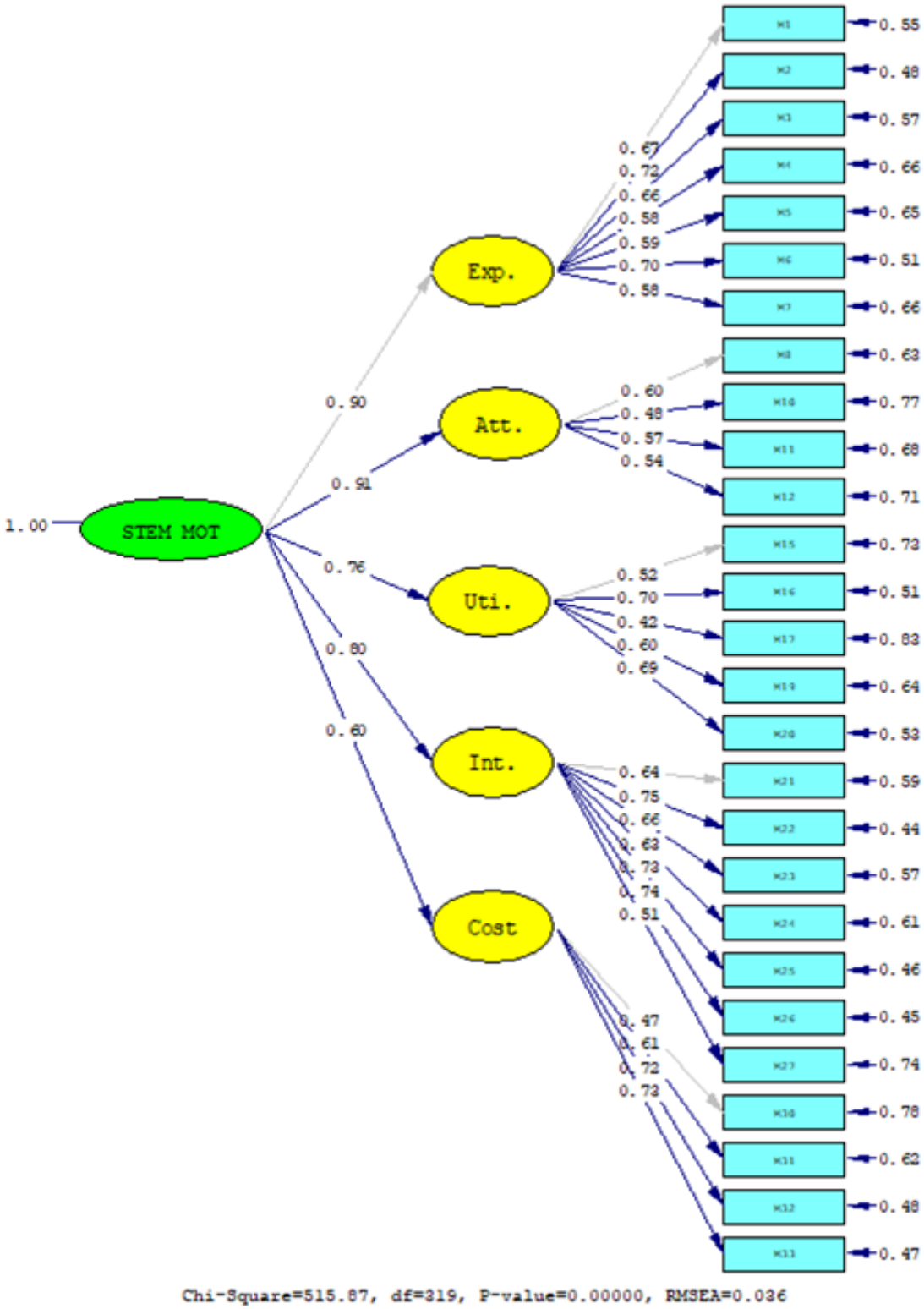
| | Mot. | Exp. | Att. | Uti. | Int. | Cost |
|---|---|---|---|---|---|---|
| Mot. | 1.00 | | | | | |
| Exp. | .90 | 1.00 | | | | |
| Att. | .91 | .82 | 1.00 | | | |
| Uti. | .76 | .69 | .61 | 1.00 | | |
| Int. | .80 | .72 | .73 | .61 | 1.00 | |
| Cost | .60 | .54 | .55 | .48 | .46 | 1.00 |

The results of a series of CFAs indicated that Model 1 does not adequately represent the factor structure, due to a high correlation between the expectancy and value factors, as well as the occurrence of a Heywood case. Additionally, Model 2 and Model 3 have similar fit indexes. However, it should be noted that Model 2 does not adequately represent the factor structure as the dimensions fail to discriminate effectively. Brown (2006) argued that if the results of CFA show strong relationships between certain factors, it is not appropriate to claim that these factors represent distinct dimensions of the structure. This finding also suggests poor discriminant validity. Additionally, in our study, a factor with a significantly higher eigenvalue compared to other factors was observed in EFA. Moreover, the high correlations between the attitude/experience and utility/attitude factors in the first-order CFA model indicate the possible presence of a second-order factor that could account for the common source of these correlations between the factors. Hence, adopting a second-order CFA model that demonstrates a comparable fit to Model 2 and incorporates a second-order factor to account for the strong correlations among the factors appeared to be a more logical approach (Iversen, et al., 2022). Based on these reasons, it was decided that utilizing Model 3 instead of Model 2 would be more suitable for this study. The second-order measurement model (Model 3) with standardized solutions is shown in Figure 3.

As seen in Figure 3 for Model 3, the Chi-square value was found to be statistically significant according to the construct validity findings of measures obtained from the 27-item scale. However, the Chi-square is sensitive to sample size (Bergh, 2015). For models with 75-200 cases, a Chi-square test is mostly a reasonable measure of fit. But for larger models (with 400 cases or more), the Chi-square is statistically significant almost always (Kenny, 2015). For that reason, examining the $\chi^2/df$ ratio is recommended (Şimşek, 2007; Waltz et al., 2010). In our study, the $\chi^2/df$ ratio for the final model was calculated as 1.61. Schermelleh-Engel et al., (2003) stated that $0 \leq \chi^2/df \leq 2$ indicates a perfect fit. Additionally, considering fit indexes described by Schermelleh-Engel et al., (2003), among the other fit indexes calculated, the GFI value displayed an acceptable fit ($.90 \leq GFI < .95$), whereas, NFI ($.95 \leq NFI \leq 1.00$), CFI ($.97 \leq CFI \leq 1.00$), and RMSEA ($0 \leq RMSEA \leq .05$) values indicate a perfect fit. Based on these findings, it can be argued that the model provides a good fit to the data. Furthermore, the factor loadings varied between .42 and .75 and the error variances were acceptable. According to Tabachnick and Fidell (2007), factor loadings greater than .71 are considered perfect, greater than .63 are very good, greater than .55 are good, greater than .45 are good/acceptable, and finally, factor

loadings greater than .32 are weak. Therefore, our findings indicate that the items represent the related factors and can make measurements with acceptable errors. Accordingly, Model 3 was decided as the valid model of the 27-item version of the STEM Motivation Scale. These findings revealed that the 27-item version of the scale can measure middle school students' STEM motivation through expectancy, intrinsic value, utility value, attainment value, and cost value dimensions.

**Figure 3.** *Second-order measurement model for STEM Motivation Scale (27-item form)*



Finally, to obtain evidence for the reliability of the measures, Cronbach's Alpha values were examined. Accordingly, Cronbach's Alpha values for the measures obtained from expectancy, utility, attainment, intrinsic value, and cost sub-scales were calculated as α=.878, α=.760, α=.700, α=. 878, and α=.729, respectively. Plus, Cronbach's Alpha of the total scale was found to be α=.921. These α values indicate an acceptable level of reliability. CFA findings and these α values were considered validity and reliability evidence for the 27-item form of the STEM Motivation Scale for middle school students.

## 4. DISCUSSION, CONCLUSION and RECOMMENDATIONS

Examination of the education period from early childhood education to college graduation is a key step for increasing the number of students interested in STEM and maintaining this interest until they receive a STEM degree (Nariman, 2021). Students' interest, persistence, and effort in STEM fields represent the whole students' achievement expectations and value perceptions for the STEM field (Açıksöz et al., 2020). To understand motivational beliefs, such as expectancy and value, that predict students' success and academic effort (Trautwein et al., 2012) and influence their persistence decisions, valid and reliable measures of these dimensions are essential. On the other hand, the lack of a reliable and practical motivation measurement tool in the literature for middle school students makes it difficult for researchers or program evaluators to determine the effectiveness of educational interventions designed to increase student motivation (Kosovich et al., 2015).

The theory introduced by Eccles et al. (1983) is composed of two main structures namely, expectancy and value. This model assumes that expectancy and value directly affect performance, persistency, and task choices (Trautwein et al., 2012). However, the sub-dimensions (utility, attainment, cost, and intrinsic value) of the expectancy and value constructs are highly correlated or loaded on a single factor mostly (Eccles & Wigfield, 1995). Thus, observing high correlations between expectancy and value as well as between value sub-dimensions is highly likely. In the current study, the results of both 33-item and 27-item scales showed that high correlations exist between factors of the 2-factor model defined based on the theory; therefore, expectancy and value constructs do not discriminate well. Consistent with our results, Trautwein et al. (2012) reported high correlations between expectancy and value beliefs.

Additionally, in the same study, Trautwein et al. (2012) found that some relationships between the sub-dimensions of value (expectancy, attainment, cost, and intrinsic value) were lower than the relationship between expectancy and value, especially, the relationship between cost and utility subdimensions was found to be low. Consistent with these, our findings indicated that the cost sub-dimension showed lower correlations compared to the relationships between other sub-dimensions. Considering other studies in which the cost sub-dimension was addressed as an empirically different construct than the expectancy and value (see Kosovich et al., 2015), it is an expected result that the cost sub-dimension did not show a higher correlation, unlike the other dimensions in our study.

EVT suggests that students' motivation for success and behaviors (preferences) are a function of their beliefs regarding their skills (expectancy) and perceived importance (value) for a specific task (Eccles et al., 1983; Wigfield et al., 2009). Considering the framework of STEM, the participation of students in STEM as well as their performance and persistence in this field can be defined as a combination of expectancy for success and perceived value in this field. Model 3, the best model according to our findings, includes all expectancy and value constructs. Moreover, this model's sufficient fit to the relevant data as well as both factor loadings and standardized unique variance estimates were good indicators of the corresponding factors can be considered evidence for the construct validity of the measures obtained from the 27-item scale. Therefore, the developed measurement tool can predict the motivation component of 5 factors based on the EVT, and the 27-item scale can yield valid measures regarding middle

school students' motivation. In addition to this, reliability results for the measures obtained from the scale showed that sub-dimensions and overall scale yield measures with an acceptable level of reliability. Based on these findings, it can be argued that the STEM Motivation Scale can address students' expectancy and the value they place on the field of STEM as a whole and can provide reliable and valid measures for middle school students' STEM motivations.

### 4.1. Use of the Scale for Research and in Teaching Environments

According to Steinmayr et al. (2019), in the limited number of studies that examined some motivational constructs as predictors of students' academic success, most of the motivational constructs predicted academic success more than intelligence, and particularly, students' ability self-concepts and task value were more powerful for predicting success. On the other hand, Areepattamannil et al. (2011) found that motivation is a predictor of academic success. However, Kulwinder Singh (2014) stated that the relationship between motivational beliefs and learning outcomes is still uncertain. In this regard, the measurement tool developed in this study can be used to explain relationships between students' motivational beliefs and academic success in STEM discipline.

Appianing and Van Eck (2018) emphasized that if one's expectations and value beliefs are high, this person is likely to stay in STEM fields, make an effort, and graduate from these fields, but otherwise, the opposite happens. Additionally, raising motivation in a specific field may help gain interest in a certain field including a future career (Hidi & Renninger, 2006). Using this measurement tool, program makers and practitioners can measure middle school students' STEM motivational beliefs, especially in formal settings and also in informal settings. Considering the constructs included in the measurement tool, the motivational dimensions of students that need to be improved can be identified and intervention practices targeting this dimension can be performed. For example, practices focusing benefits of a task or discipline can be carried out for students who were identified with lower utility value, on the other hand, practices improving self-efficacy beliefs can be implemented for students who consider themselves inadequate (those with lower expectancy) for an activity or discipline. In this regard, this measurement tool can be a guide for determining strategies aiming to improve students' STEM motivation or designing curricula according to these needs.

Furthermore, aiming for student motivation only in a certain period might be insufficient to meet future STEM workforce needs. Although our study was carried out for middle school STEM fields, we know that students may leave STEM in the further educational stages. This is why we consider validating this measurement tool by implementing it in different education levels (high school, university) important. Moreover, this measurement tool, which we believe is important in terms of its potential contribution to further research and intervention strategies, was validated by the data collected from a specific socio-cultural population and in an urban region in Turkey. Accordingly, validating this measurement tool with populations of different languages and cultures would contribute to the validity and reliability studies of the scale.

### 4.2. Conclusion

Since students' preference, persistence, and performance in STEM fields, whose importance is constantly rising in today's world, are partly shaped by students' motivation, more studies are needed to understand motivation dynamics. This measurement tool, which can make valid and reliable measurements, allows for determining motivational beliefs within the expectancy-value concept that can be targeted to encourage students' interest in STEM fields as well as help design interventions for these structure(s) and evaluate the effectiveness of these interventions.

### Acknowledgments

The authors would like to thank Gazi University Academic Writing Application and Research Center for proofreading the article.

## 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 Üniversitesi, Ölçme Değerlendirme Etik Alt Çalışma Grubu, 17.12.2020-E.135741.

## Contribution of Authors

All authors have equally contributed to all section of this study. The authors read and approved the final manuscript.

## Orcid

Arif Açıksöz https://orcid.org/0000-0002-6770-3777
İlbilge Dökme https://orcid.org/0000-0003-0227-6193
Emine Önen https://orcid.org/0000-0002-0398-3191

## REFERENCES

Açıksöz, A., Özkan, Y., & Dökme, İ. (2020). Adaptation of the STEM value-expectancy assessment scale to Turkish culture. *International Journal of Assessment Tools in Education, 7*(2), 177-190. https://doi.org/10.21449/ijate.723408

Appianing J., & van Eck, R.N. (2018). Development and validation of the Value-Expectancy STEM Assessment Scale for students in higher education. *International Journal of STEM Education, 5*(24), 1-16. https://doi.org/10.1186/s40594-018-0121-8

Areepattamannil, S., Freeman, J.G., & Klinger, D.A. (2010). Influence of motivation, self-beliefs, and instructional practices on science achievement of adolescents in Canada. *Social Psychology of Education, 14*(2), *233–259.* https://doi.org/10.1007/s11218-010-9144-9

Ball, C., Huang, K.T., Cotten, S.R., & Rikard, R.V. (2017). Pressurizing the STEM pipeline: An expectancy-value theory analysis of youths' STEM attitudes. *Journal of Science Education and Technology, 26*(4), 372-382. https://doi.org/10.1007/s10956-017-9685-1

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W H Freeman/Times Books/ Henry Holt & Co.

Barron, K.E., & Hulleman, C.S. (2015). Expectancy-Value-Cost model of motivation. In J.D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences,* (2nd Ed., Vol.8, pp. 503-509). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.26099-6

Barutcu, T. (2017). *Beklenti-değer temelli öğretimde yazma becerileri ve motivasyon ilişkisi [The relation between writing skills and motivation in teaching based upon the expectancy-value]* [Doctoral dissertation, Gazi University].

Beerenwinkel, A., & von Arx, M. (2017). Constructivism in practice: An exploratory study of teaching patterns and student motivation in physics classrooms in Finland, Germany and Switzerland. *Research in Science Education, 47*, 237-255. http://doi.org/10.1007/s11165-015-9497-3

Bergh, D. (2015). Chi-Squared test of fit and sample size: A comparison between a random sample approach and a chi-square value adjustment method. *Journal of Applied Measurement, 16*(2), 204–217.

Bøe, M.V., Henriksen, E.K., Lyons, T., & Schreiner, C. (2011). Participation in science and technology: Young people's achievement-related choices in late modern societies. *Studies in Science Education, 47*(1), 1-36. https://doi.org/10.1080/03057267.2011.549621

Breiner, J., Harkness, S., Johnson, C., & Koehler, C. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics, 112*(1), 3–11. https://doi.org/10.1111/j.1949-8594.2011.00109.x

Brown, T.A. (2006). *Confirmatory factor analysis for applied research.* The Guilford Press.

Brown, P.L., Concannon, J.P., Marx, D., Donaldson, C., & Black, A. (2016). An examination of middle school students' STEM self-efficacy, interests and perceptions. *Journal of STEM Education: Innovations and Research*, *17*(3), 27-38

Bybee, R.W. (2010). What is STEM education?. *Science,* 329(5995), 996.

Chen, J.A., & Dede, C.J. (2011). Youth STEM motivation: Immersive technologies to engage and empower underrepresented students. In *Conference proceeding TEST Learning Resource Center at EDC*. Retrieved February 15, 2022 from https://publish.wm.edu/workingpapers/1

Corlu, M.S., Capraro, R.M., & Capraro, M.M. (2014). Introducing STEM education: Implications for educating our teachers in the age of innovation. *Education and Science, 39*(171), 74-85.

Christensen, R., & Knezek, G. (2017). Relationship of middle school student STEM interest to career intent. *Journal of Education in Science, Environment and Health (JESEH), 3*(1), 1-13. https://doi.org/10.21891/jeseh.275649

Dabney, K.P., Tai, R.H., Almarode, J.T., Miller-Friedmann, J.L., Sonnert, G., Sadler, P.M., & Hazari, Z. (2012). Out-of-school time science activities and their association with career interest in STEM. *International Journal of Science Education, Part B, 2*(1), 63–79. https://doi.org/10.1080/21548455.2011.629455

DeVellis, R.F. (2003). *Scale development theory and applications* (2nd Ed.). SAGE Publication.

Dotterer, A.M. (2022). Parent involvement, expectancy values, and STEM outcomes among underrepresented adolescents. *Social Psychology of Education*, 25, 113-127. https://doi.org/10.1007/s11218-021-09677-0

Eccles, J.S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A.J. Elliot & C.S. Dweck (Eds.), *Handbook of Competence and Motivation* (pp. 105-121). Guilford Press.

Eccles, J.S. (2011). Gendered educational and occupational choices: Applying the Eccles et al. model of achievement-related choices. *International Journal of Behavioral Development*, *35*(3) 195–201. https://doi.org/10.1177/0165025411398185

Eccles-Parsons, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J.T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-146). W.H. Freeman.

Eccles, J.S., & Harold, R.D. (1991). Gender differences in sport involvement: Applying the Eccles' expectancy-value model. *Journal of Applied Sport Psychology, 3,* 7-35. https://doi.org/10.1080/10413209108406432

Eccles, J.S., Roeser, R., Wigfield, A., & Freedman-Doan, C. (1999). Academic and motivational pathways through middle childhood. In L. Balter & C.S. Tamis-LeMonda (Eds.), *Child psychology: A handbook of contemporary issues* (pp. 287-317). Psychology Press.

Eccles, J.S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin, 21*(3), 215–225.

Eccles, J.S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109-132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eccles, J.S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61,* 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

English, L.D. (2017) Advancing elementary and middle school STEM education. *International Journal of Education in Mathematics, Science and Technology, 15*(1), 5-24. https://doi.org/10.1007/s10763-017-9802-x

French, A.M., Else-Quest, N.M., Asher, M., Thoman, D.B., Smith, J.L., Hyde, J.S., & Harackiewicz, J.M. (2023). An intersectional application of expectancy-value theory in an undergraduate chemistry course. *Psychology of Women Quarterly*, *47*(3), 299-319.

Glynn, S.M., Brickman, P., Armstrong, N., & Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and non-science majors. *Journal of Research in Science Teaching*, *48*(10), 1159-1176. https://doi.org/10.1002/tea.20442

Gok, T. (2021). The development of the STEM (science, technology, engineering, and mathematics) attitude and motivation survey towards secondary school students. *International Journal of Cognitive Research in Science, Engineering and Education*, *9*(1), 105-119. https://doi.org/10.23947/2334-8496-2021-9-1-105-119

Gonzalez, H.B., & Kuenzi, J.J. (2012). *Science, technology, engineering, and mathematics (STEM) education: A primer.* Congressional Research Service, Library of Congress. Washington, DC. http://www.fas.org/sgp/crs/misc/R42642.pdf

Gråstén, A. (2016). Children's expectancy beliefs and subjective task values through two years of school-based program and associated links to physical education enjoyment and physical activity. *Journal of Sport and Health Science*, *5*(4), 500-509.

Griffith, A.L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review*, *29*(6), 911-922. https://doi.org/10.1016/j.econedurev.2010.06.010

Hermans, S., Gijsen, M., Mombaers, T., & van Petegen, P. (2022). Gendered patterns in students' motivation profiles regarding iSTEM and STEM test scores: A cluster analysis. *International Journal of STEM Education*, *9*, 67. https://doi.org/10.1186/s40594-022-00379-3

Hidi, S., & Renninger, K.A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*, 111–127. https://doi.org/10.1207/s15326985ep4102_4

Hinton Jr, A.O., Termini, C.M., Spencer, E.C., Rutaganira, F.U., Chery, D., Roby, R., ... & Palavicino-Maggio, C. B. (2020). Patching the leaks: revitalizing and reimagining the STEM pipeline. *Cell*, *183*(3), 568-575. https://doi.org/10.1016/j.cell.2020.09.029

Hossain, M., & Robinson, M.G. (2012). How to motivate US students to pursue STEM (science, technology, engineering and mathematics) careers. *US-China Educ Rev A*, *4*, 442–451.

Iversen, M.M., Norekvål, T.M., Oterhals, K., Fadnes, L.T., Mæland, S., Pakpour, A.H., & Breivik, K. (2022). Psychometric properties of the Norwegian version of the fear of COVID-19 Scale. *International Journal of Mental Health and Addiction*, 1-19. https://doi.org/10.1007/s11469-020-00454-2

Jones, B.D. (2009). Motivating students to engage in learning: The MUSIC Model of Academic Motivation. *International Journal of Teaching and Learning in Higher Education*, *21*(2), 272-285.

Jones, B.D. (2018). *Motivating students by design: Practical strategies for professors* (2nd Ed.). CreateSpace.

Jones, B.D. (2012/2022, November). *User guide for assessing the components of the MUSIC Model of Motivation.* http://www.theMUSICmodel.com

Joseph, C.H., Anikelechi, I.G., & Marumo, P. (2019). Academic motivation of school going adolescents: Gender and age difference. *Gender and Behaviour*, *17*(1), 12306-12315.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151. https://doi.org/10.1177/001316446002000116

Kenny, D. (2015). *Measuring model fit*. Retrieved March 10, 2023, from http://davidakenny.net/cm/fit.htm

Kızılay, E., Yamak, H., & Kavak, N. (2019). Motivation scale for STEM fields. *Journal of Computer and Education Research*, *7*(14), 540-557. https://doi.org/10.18009/jcer.617514

Kosovich, J.J., Hulleman, C.S., Barron, K.E., & Getty, S. (2015). A practical measure of student motivation: Establishing validity evidence for the Expectancy-Value-Cost Scale in

middle school. *The Journal of Early Adolescence*, *35*(5-6), 790-816. https://doi.org/10.1177/0272431614556890

Kuenzi, J. (2008). *Science, technology, engineering, and mathematics (STEM) education: Background, federal policy and legislative action.* Congressional Research Service Reports. Retrieved February 10, 2022 from http://digitalcommons.unl.edu/crsdocs/35/

Kulwinder Singh, K. (2014). Motivational beliefs and academic achievement of university students. *IOSR Journal of Research & Method in Education, 4(1),* 1-3.

Kuo, H.C., Tseng, Y.C., & Yang, Y.T.C. (2019). Promoting college student's learning motivation and creativity through a STEM interdisciplinary PBL human-computer interaction system design and development course. *Thinking Skills and Creativity, 31*, 1-10. https://doi.org/10.1016/j.tsc.2018.09.001

Knezek, G., Christensen, R., Tyler-Wood, T., & Periathiruvadi, S. (2013). Impact of environmental power monitoring activities on middle school student perceptions of STEM. *Science Education International*, *24*(1), 98-123.

Langdon, D., McKittrick, G., Beede, D., Khan, B., & Doms, M. (2011). *STEM: Good jobs now and for the future* (ESA Issue Brief 03-11). U.S. Department of Commerce. Economics and Statistics Administration. https://files.eric.ed.gov/fulltext/ED522129.pdf

Loh, E.K. (2019). What we know about expectancy-value theory, and how it helps to design a sustained motivating learning environment. *System*, *86*, 102119.

Luo, T., So, W.W.M., Wan, Z.H., & Li, W.C. (2021). STEM stereotypes predict students' STEM career interest via self-efficacy and outcome expectations. *International Journal of STEM Education*, *8*(1), 1-13. https://doi.org/10.1186/s40594-021-00295-y

Luo, T., Wang, J., Liu, X., & Zhou, J. (2019). Development and application of a scale to measure students' STEM continuing motivation. *International Journal of Science Education*, *41*(14), 1885-1904. https://doi.org/10.1080/09500693.2019.1647472

Meyer, J., Fleckenstein, J., & Köller, O. (2019). Expectancy value interactions and academic achievement: Differential relationships with achievement measures. *Contemporary Educational Psychology*, *58*, 58–74. https://doi.org/10.1016/j.cedpsych.2019.01.006

Moreno, N.P., Tharp, B.Z., Vogt, G., Newell, A.D., & Burnett, C.A. (2016). Preparing students for middle school through after-school STEM activities. *Journal of Science Education and Technology*, *25*, 889-897. https://doi.org/10.1007/s10956-016-9643-3

Nariman, N. (2021). How does an industry-aligned technology-rich problem-based learning (PBL) model influence low-income and native Hawaiian student's STEM career interest?. *Journal of Problem Based Learning in Higher Education*, *9*(1), 150-178. https://doi.org/10.5278/ojs.jpblhe.v9i1.6367

National Research Council [NRC] (2011). *Successful K-12 STEM education. Identify effective approaches in science, technology, engineering and mathematics*. The National Academies Press. https://www.ltrr.arizona.edu/webhome/sheppard/TUSD/NRC2011.pdf

National Science and Technology Council [NSTC] (2018). *Charting a course for success. America's strategy for STEM education.* https://www.whitehouse.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf

PCAST (2010). *Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Math (STEM) Education for America's Future Executive Report.* National Science Foundation. https://nsf.gov/attachments/117803/public/2a--Prepare_and_Inspire--PCAST.pdf

Partridge, J., Brustad, R., & Stellino, M.B. (2013). Theoretical perspectives: Eccles's expectancy value theory. *Advances in Sport Psychology*, *3*, 269-292.

Perez, T., Wormington, S.V., Barger, M.M., Schwartz-Bloom, R.D., Lee, Y., & Linnenbrink-Garcia, L. (2019). Science expectancy, value, and cost profiles and their proximal and distal relations to undergraduate science, technology, engineering, and math persistence. *Science Education*, *103*, 264-286. https://doi.org/10.1002/sn.21490

Petri, H.L., & Govern, J.M. (2013). *Motivation: Theory, research, and application* (6th ed.). Wadsworth Cenpage Learning.

Razali, F. (2021). Exploring crucial factors of an interest in STEM career model among secondary school students. *International Journal of Instruction*, *14*(2), 385-404. https://doi.org/10.29333/iji.2021.14222a

Robnett., R.D., & Leaper, C. (2012). Friendship groups, personal motivation, and gender in relation to high school students' STEM career. *Interest Journal of Research on Adolescence, 23*(4), 652–664. https://doi.org/10.1111/jora.12013

Romine, W.L., & Sadler, T.D. (2016). Measuring changes in interest in science and technology at the college level in response to two instructional interventions. *Research in Science Education. 46*(3), 309-327.

Rosenzweig, E.Q., & Wigfield, A. (2016). STEM motivation interventions for adolescents: A promising start, but further to go. *Educational Psychologist, 51*(2), 146-163.

Rosenzweig, E.Q., Wigfield, A., & Eccles J.S. (2019). Expectancy – value theory and its relevance for student motivation and learning. In K.A. Renninger & S.E. Hidi (Eds.), *The Cambridge handbook of motivation and learning* (pp. 617-644). Cambridge University Press. https://doi.org/10.1017/9781316823279.026

Rozek, C.S., Svoboda, R.C., Harackiewicz, J.M., Hulleman, C.S., & Hyde, J.S. (2017). Utility-value intervention with parents' increases students' STEM preparation and career pursuit. *Proceedings of the National Academy of Sciences*, *114*(5), 909-914. https://doi.org/10.1073/pnas.1607386114

Sadler, P.M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of stem career interest in high school: A gender study. *Science Education, 96*(3), 411–427. https://doi.org/10.1002/sce.21007

Sanders, M. (2009). STEM, STEM education, STEMmania. *The Technology Teacher*, *68*(4), 20-26.

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, 8*(2), 23–74.

Schoenherr, J. (2024). Personalizing real-world problems: Posing own problems increases self-efficacy expectations, intrinsic value, attainment value, and utility value. *British Journal of Educational Psychology*, 1-18.

Shin, D.D., Lee, M., Ha, J.E., Park, J.H., Ahn, H.S., Son, E., Chung, Y., & Bong, M. (2019). Science for all: Boosting the science motivation of elementary school students with utility value intervention. *Learning and Instruction.* 60, 104-116. https://doi.org/10.1016/j.learninstruc.2018.12.003

Steinmayr, R., Weidinger, A.F., Schwinger, M., & Spinath, B. (2019). The importance of students' motivation for their academic achievement: Replicating and extending previous findings. *Frontiers in Psychology*, *10*, 1730. https://doi.org/10.3389/fpsyg.2019.01730

Şimşek, Ö.F. (2007). *Yapısal eşitlik modellemesine giriş: Temel ilkeler ve LISREL uygulamaları [Introduction to structural equation modeling: Basic principles and LISREL applications]*. Ekinoks Yayınları, Ankara.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics.* Pearson Education, Inc.

Tai, R., Liu, C., Maltese, A., & Fan, X. (2006). Planning early for careers in science. *Science, 312*(5777), 1143- 1144. https://doi.org/10.1126/science.1128690

Trautwein, U., Marsh, H.W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology, 104*(3), 763–777. https://doi.org/10.1037/a0027470

Valenzuela, J., Nieto, A.M., & Saiz, C. (2011). Critical thinking motivational scale: a contribution to the study of relationship between critical thinking and motivation. *Journal*

*of Research in Educational Psychology*, *9*(2), 823-848. https://doi.org/10.25115/ejrep.v9i24.1475

Van den Hurk, A., Meelissen, M., & Van Langen, A. (2019). Interventions in education to prevent STEM pipeline leakage. *International Journal of Science Education, 41*(2), 150-164. https://doi.org/10.1080/09500693.2018.1540897

Walters, S., Santana, C., Zastavker, Y.V., Dillon, A., Stolk, J.D., & Gross, M.D. (2016). Students' motivational attitudes in introductory STEM courses: The relationship between assessment and externalization. In *2016 IEEE Frontiers in Education Conference (FIE)* (pp. 1-4). IEEE.

Waltz, C.F., Strickland, O., & Lenz, E.R. (2010). *Measurement in nursing and health research.* Springer Publishing Company, New York.

Wang, M., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review*, *33*(4), 1-37. https://doi.org/10.1016/j.dr.2013.08.001

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal, 50*(5), 1081-1121. https://doi.org/10.3102/0002831213488622

Wentzel, K.R., & Wigfield, A. (1998). Academic and social motivational influences on students' academic performance. *Educational Psychology Review*, *10*(2), 155-175. https://doi.org/10.1023/A:1022137619834

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49-78. https://doi.org/10.1007/BF02209024

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*, 1–35. https://doi.org/10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J.S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*, 265-310. https://doi.org/10.1016/0273-2297(92)90011-P

Wigfield, A., & Eccles, J.S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68-81. https://doi.org/10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J.S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J.S. Eccles (Eds.). *Development of achievement motivation.* Academic Press. https://doi.org/10.1016/B978-012750053-9/50006-1

Wigfield, A., & Gladstone, J.R. (2019). What does expectancy-value theory have to say about motivation and achievement in times of change and uncertainty?. In E. N. Gonida & M. S. Lemos (Eds.), *Motivation in Education at a Time of Global Change* (pp. 15–32). Emerald. https://doi.org/10.1108/s0749-742320190000020002

Wigfield, A., Rosenzweig, E.Q., & Eccles, J.S. (2017). Achievement values: Interactions, interventions, and future direction. In A.J. Elliot, C.S. Dweck & D.S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (116–134). The Guilford Press.

Wigfield, A., Tonks, S., & Klauda, S.L. (2009). Expectancy-value theory. In K.R. Wentzel & A. Wigfield (Eds.), *Educational psychology handbook series. Handbook of motivation at school* (pp. 55–75). Routledge/Taylor & Francis Group.

Williams, K.C., & Williams, C.C. (2011). Five key ingredients for improving student motivation. *Research in Higher Education Journal, 12*, 1-23.

Wu, F., & Fan, W. (2017). Academic procrastination in linking motivation and achievement-related behaviours: A perspective of expectancy value theory. *Educational Psychology, 37*(6), 695-711.

Xiang, P., Mcbride, R., Guan, J., & Solmon, M. (2003). Children's motivation in elementary physical education: An expectancy-value model of achievement choice. *Research Quarterly for Exercise and Sport, 74*, 25-35. https://doi.org/10.1080/02701367.2003.10609061

Yahaya, J., Fadzli, S., Deraman, A., Yahaya, N.Z., Halim, L., Rais, I.A.I., & Ibrahim, S.R.A. (2022). RInK: Environmental virtual interactive based education and learning model for STEM motivation. *Educ Inf Technol, 27*, 4771–4791. https://doi.org/10.1007/s10639-021-10794-8

Yurt, E. (2016). Examination of task values and expectancy beliefs of middle school students towards mathematics. *International Online Journal of Educational Sciences, 8*(1), 200-215.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 721–732
https://doi.org/10.21449/ijate.1445219
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Imaginary latent variables: Empirical testing for detecting deficiency in reflective measures

**Marco Vassallo**[1*]

[1]CREA, Research Centre for Agricultural Policies and Bioeconomy, Rome, Italy

**ARTICLE HISTORY**
*Received: Feb. 29, 2024*
*Accepted: July 21, 2024*

**Keywords:**
Latent variable models,
Structural Equation Modeling,
Imaginary and complex numbers,
Path analysis,
Psychometrics.

**Abstract:** Imaginary latent variables are variables with negative variances and have been used to implement constraints in measurement models. This article aimed to advance this practice and rationalize the imaginary latent variables as a method to detect possible latent deficiencies in measurement models. This rationale is based on the theory of complex numbers used in the measurement process of common factor model–based structural equation modeling. Modeling an imaginary latent variable produces a potential deficiency within its relative reflective measures through a considerable reduction in common variance indicating the most affected indicator(s).

## 1. INTRODUCTION

Rindskopf (1984, p. 38) first defined imaginary latent variables as: "… variables with negative variances, or, equivalently, variables with positive variance but whose influence on other variables is represented by an imaginary rather than a real number.'' These variables are of no interest themselves, but only exist to implement the constraints." Considering the first situation, in which an imaginary latent variable has a negative variance, what might it mean in applied psychological and/or educational measurement? Above and beyond of implementing constraints? Might it be useful for detecting potential latent variable deficiency?

Rindskopf (1984) described the use of imaginary latent variables by recalling Bentler and Lee's (1983) work where imaginary latent variables were used by fixing the variances to –1 to permit a measurement model having factors with the same variance as 1: a computational detracting strategy to allow the covariance matrix being able to run the correlational structure. However, this empirical exercise did not reveal the usefulness of the imaginary latent variable unless it was used as a constraint to produce equality restrictions in linear structural models. In my view, constraining a latent variable to be imaginary is not limited to a computational way to implement constraints in measurement models; however, it has potential conceptual

*CONTACT: Marco VASSALLO ✉ marco.vassallo@crea.gov.it CREA, Research Centre for Agricultural Policies and Bioeconomy



**e-ISSN: 2148-7456**

implications in the underlying measures that, as will be explained in the next section, have ground in the field of the imaginary complex numbers.

Essentially, and to be as reasonable as possible, whenever a latent variable is considered as imaginary, with its negative variance, a researcher postulates a sort of "what if" scenario concerning a potential deficiency of that latent concept in a specific context. That is to say, this imaginary interrogation may want to test what could happen to a latent concept if it has been affected by some causes that have triggered its absence. Consequently, this deficiency will be spread throughout those observed measures that are a reflection, manifestation, and an effect of that latent concept. These observed measures (i.e., the well-known reflective indicators of a latent variable) under imaginary interrogation can determine which aspects of that latent concept are more affected by this potential deficiency/absence.

In this respect, the aim of this article was to propose a simple empirical test based on constraining latent variables to become imaginary and thus verifying what could happen to their reflective measures if they are affected by a potential deficiency in a measurement model and hence within a context of application.

The remainder of this paper is structured as follows. The next section presents the conceptual foundations of this imaginary latent process in a measurement model. Successively, the following section presents a computational demonstration on Schwartz' (1992) human values taxonomy applied to an Italian sample. Finally, a short discussion with limitations and future perspectives concludes this work.

### 1.1. The Parallelism Between The Measurement Process of Common Factor Models and The Rationale of Imaginary Complex Numbers

The logic and rationale of the classic measurement process, taken from the classic measurement process based on the classical test theory (Lord & Novick, 1968) of true and error scores, postulates that any measure $x_i$, even the one obtained with the most sophisticated procedures, is affected by a measurement error $e_i$ (a nonsystematic but normally distributed with zero mean and nonzero variance); therefore, this measure is functional/dependent on the true measure $t_i$ (which may be latent in nature and thereby unknown) and the measurement error itself:

$$x_i = t_i + e_i \tag{1}$$

As a logical computational consequence, the true measure is indeed the expected value of the initial measures and is not related to the measurement error:

$$E(x_i) = t_i \tag{2}$$

$$Cov(t_i, e_i) = 0 \tag{3}$$

According to Equations 1 and 3, a researcher may have a set of observed measures $x_i$ with variances $\sigma^2_{xi}$ that can be decomposed of another set of true measures with latent true error–free variable variances $\sigma^2_{ti}$ and a set of measurement errors with variances $\sigma^2_{ei}$:

$$\sigma^2_{xi} = \sigma^2_{ti} + \sigma^2_{ei} \tag{4}$$

$$\rho = \sigma^2_{ti}/\sigma^2_{xi} \tag{5}$$

Equation 4 depicts the famous definition of reliability[†] ρ (5) of the classic measurement process where a true value is a value free of measurement error. This true value is indeed a value that is still unknown and requires a set of observed measures to be revealed as precisely as possible by partial-out measurement errors from the common values.

In connection therewith, we know the common factor model theory of Thurstone (1947), which

[†]"Reliability is the ratio of true score's variance to the observed variable's variance" (Bollen, 1989, p.208).

constitutes the key to factor analysis, that each set of observed variables may be written, or better decomposed of, as a linear function of that part of common shared variance and that part that is unique in each observed itself. These two concepts of common shared variance and unique variance represent what have been above formalized with the expression (4) where $\sigma^2_{ti}$ is the common shared variance needed to reflect the manifestation of a common latent factor (i.e., the true value to be sought), whereas $\sigma^2_{ei}$ is the unique variance that embodies the following: (a) the part of the observed variance that each observed variable does not share with the observed variances of the other observed variables and thus not useful to manifest the true value and (b) the random error owing to the measurement process.

Hence, by combining the classical test theory of measurement process with a typical confirmatory factor analysis (CFA) model (Bollen, 1989; Jöreskog, 1966), a type[‡] of common factor model where the relations between measures and factors are a priori specified, Equation 1 can be explicated in a system of simple linear regression equations as follows:

$$x_i = \tau_i + \lambda_i \xi + \delta_I \quad (6)$$

where $x_i$ is a set of observed variables (i = 1, …, n), $\xi$ is a hypothetical common latent factor, $\lambda_i$ is the factor loading or regression slope, $\tau_i$ is the intercept, and $\delta_i$ is the measurement error. The difference between Equation 6 and a typical regression equation is that the independent variable is the latent factor $\xi$ and the criterion is constituted by multiple observed variables $x_i$. Therefore, it does mean that the latent concept $\xi$ is trying to explain, and summarize, all those observed variables $x_i$, and the magnitude of how much the latent factor can do that is owing to the regression slopes or factor loadings $\lambda_i$ associated with each $x_i$ . The magnitude of what was not captured by the latent factor is $\delta_i$, which is an error in this sort of interpolation process. This error has an expected value E ($\delta_i$) = 0 and Cov ($\xi$;$\delta_i$) = 0.

Equation 6 estimates parameters $\tau_i$, $\lambda_i$, and $\delta_i$ using all the information of the observed measures $x_i$ that constitute all the sources of covariation of $x_i$: the variances and covariances of each involved $x_i$. This leads to the fundamentals of the structural equation model applied to measured variable and latent variables path analysis (Bollen, 1989): decomposition of observed variances and covariances (i.e., the matrix $\Sigma_{xx}$) into the model-implied parameters (i.e., the model-implied matrix $\Sigma$ ($\theta$)):

$$\Sigma = \Sigma[\theta] \quad (7)$$

If a researcher can write the system of Equation 7 he/she can list all the necessary parameters of the model (6).

For an example with two-latent factors $\xi_1$ and $\xi_2$ and four measures ($x_1$, $x_2$, $x_3$, $x_4$) as depicted in Figure 1[§], it is possible to rewrite the covariance matrix of the four measures following the system of Equations 6, as shown in Table 1.

[‡]The other type of common factor model is the famous explorative factor analysis (EFA) where the relations between measures and factors are not a priori specified. EFA and CFA can partial out common variance from unique variance. However, the former assumes measurement error at random; hence, it cannot be modeled while the latter may assume measurement error at random, or not, and thus it can be modeled (Brown, 2006; Fabricar et al., 1999).

[§]The model in Figure 1 is not identified, and it requires to fix one of the $\lambda_i$ to 1 for each latent factor. As soon as this identification is done, relative decomposition Table 1 will be simplified accordingly, and the imaginary process will involve only the other not fixed $\lambda_i$. However, for a better understanding of the process, I did not indicate either in Figure 1 or Table 1 that the $\lambda_i$ needs to be equal to 1 to trigger the idea that all the $\lambda_i$ must be involved into the imaginary process alternatively as described in the results section.

**Figure 1.** *Path diagram of two common factors with four measures.*

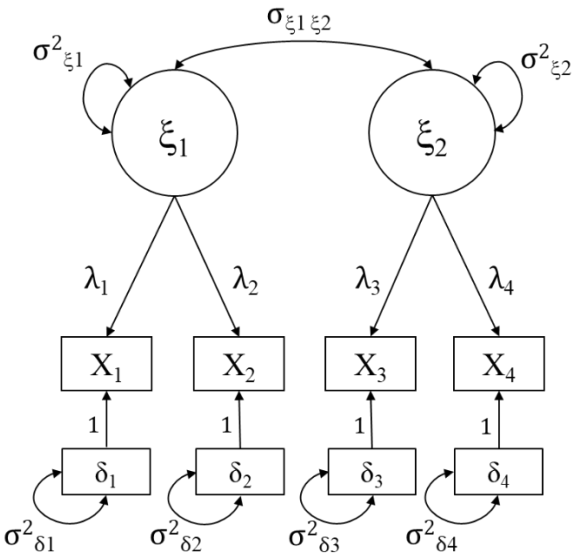



**Table 1.** *Decomposition table of structural parameters of two common factor models with four measures (adapted from Hancock et al., 2009).*

| | | Unknown parameters | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| info | decomposition | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\sigma^2_{\xi1}$ | $\sigma_{\xi1\xi2}$ | $\sigma^2_{\xi2}$ | $\sigma^2_{\delta1}$ | $\sigma^2_{\delta2}$ | $\sigma^2_{\delta3}$ | $\sigma^2_{\delta4}$ |
| $\sigma^2_{x1}$ | $\lambda_1^2\sigma^2_{\xi1} + \sigma^2_{\delta1}$ | √ | | | | √ | | | √ | | | |
| $\sigma^2_{x2}$ | $\lambda_2^2\sigma^2_{\xi1} + \sigma^2_{\delta2}$ | | √ | | | √ | | | | √ | | |
| $\sigma^2_{x3}$ | $\lambda_3^2\sigma^2_{\xi2} + \sigma^2_{\delta3}$ | | | √ | | | | √ | | | √ | |
| $\sigma^2_{x4}$ | $\lambda_4^2\sigma^2_{\xi2} + \sigma^2_{\delta4}$ | | | | √ | | | √ | | | | √ |
| $\sigma_{x1x2}$ | $\lambda_1\lambda_2\sigma^2_{\xi1}$ | √ | √ | | | √ | | | | | | |
| $\sigma_{x1x3}$ | $\lambda_1\lambda_3\sigma_{\xi1\xi2}$ | √ | | √ | | | √ | | | | | |
| $\sigma_{x1x4}$ | $\lambda_1\lambda_4\sigma_{\xi1\xi2}$ | √ | | | √ | | √ | | | | | |
| $\sigma_{x2x3}$ | $\lambda_2\lambda_3\sigma_{\xi1\xi2}$ | | √ | √ | | | √ | | | | | |
| $\sigma_{x2x4}$ | $\lambda_2\lambda_4\sigma_{\xi1\xi2}$ | | √ | | √ | | √ | | | | | |
| $\sigma_{x3x4}$ | $\lambda_3\lambda_4\sigma^2_{\xi2}$ | | | √ | √ | | | √ | | | | |

Reading the table horizontally indicates how many and which piece of information we need to estimate the unknown parameters (Hancock et al., 2009). On the contrary, by reading the table vertically, we are aware of which decomposition expression is directly involved in the estimation of that particular parameter (Hancock et al., 2009). The checkmarks indicate the combinations. It is noteworthy that to estimate the latent variances $\sigma^2_{\xi1}$ and $\sigma^2_{\xi2}$, we require all the information available in the observed measures as expected. Furthermore, the latent variances are functions of all other parameters because they are involved in almost all the decomposition expressions, although unevenly.

Considering the abovementioned, and recalling the theory of imaginary and complex numbers, we acknowledge that an imaginary number is $i^2 = -1$ (or $i = \sqrt{-1}$), and thus a complex number is the sum of a real number x with an imaginary part i (i.e., x + i when the weight of i is 1); on the contrary, a latent variable (LV) is imaginary if its variance (var) is negative (i.e., var (LV) = −1), and thus looking again at decomposition Table 1 for the latent variances, the following

expression for the measure $x_1$ (the same for the other three left) can be written as

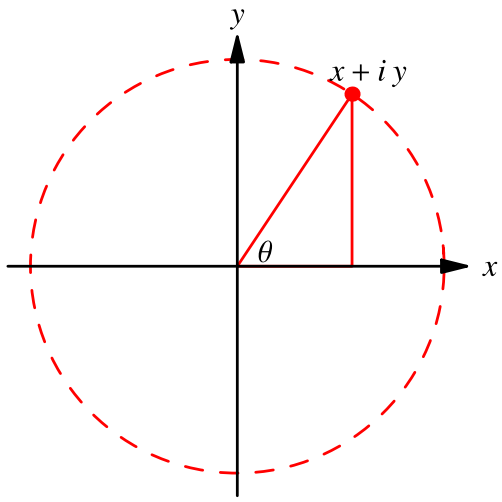
$$\sigma^2_{x1} = \lambda^2_1(-1) + \sigma^2_{\delta 1} \quad (8)$$

$$\sigma^2_{x1} - \sigma^2_{\delta 1} = \lambda^2_1(-1) \quad (9)$$

From Equation 9, we know that when an imaginary latent is postulated, the relative common variance $\lambda^2_i$ is negative, which can occur when the unique variances are high. From decomposition Table 1 and Equation 9, it is straightforward noticing how this process involves all the four measures.

This intuition becomes a deduction while referring to the properties of imaginary numbers and thus to the well-known complex number geometrical representation of the Argand diagram (Weisstein, 2023), as shown in Figure 2, where the imaginary part *iy* is on the vertical axis, whereas the real numbers *x* are on the horizontal axis.

**Figure 2.** *The Argand diagram (Weisstein, 2023).*




The logic of the circle is as follows: The more the real number *x* increases, the more the imaginary part *iy* decreases. By translating this rationale to the case of imaginary LVs, the same logic can be applied to its reflective measures. To measure $x_1$ in Equations 8 and 9, the more the unique variance $\sigma^2_{\delta 1}$ (i.e., the real number x in Figure 2) increases, the more the common variance $i\lambda^2_1$ decreases (i.e., the imaginary part *iy* in Figure 2): This explains the deficiency in items while posing var (LV) to −1, to let it imaginary.

Therefore, it seems reasonable to assume that an imaginary LV is not a proper variable that does not exist because its variance is not zero but equal to a number, although imaginary. Hence, constraining a latent factor to have a negative variance seems to hypothesize *what* could happen *if*, for some reason, there was a deficiency in that factor within its measurement model. Consequently, this deficiency spreads out within its reflective measures, most precisely affecting the common variances (i.e., factor loadings). This can pragmatically indicate which items might be more affected by a potential latent deficiency and suggest which latent aspects (i.e., measures) a specific sample of respondents may be deficient in. The estimation process of the system (7) for the two-latent model in Figure 1 with the imaginary testing with Equation 8 (i.e., by constraining the latent variance $\sigma^2_{\xi 1}$ to –1) will yield to new factor loadings values affected by the imaginary constraint. Furthermore, in the decomposition properties in Table 1, even the estimated latent covariance $\sigma_{\xi 1 \xi 2}$ will be affected by the factor loading modifications, and thus the deficiency in the latent $\xi_1$ will possibly modify the relation with the other latent $\xi_2$ as well.

## 2. METHOD AND METHODS: An example of imaginary latent process

An empirical example of the proposed imaginary latent process will be conducted from the European Social Survey (ESS) (ESS Round 10: European Social Survey, 2022) Italian data of the latest round 10 (ESS Round 10: European Social Survey Round 10 Data, 2020). The ESS is a biennial cross-national survey organized by the European Research Infrastructure Consortium to collect data on the attitudes, values, beliefs, and many behavioral patterns of European countries citizens.

The Schwartz human values section H of the ESS questionnaire (ESS Round 10: European Social Survey, 2022) will be used to select items relative to the two domains of Universalism and Benevolence.

Universalism

(1) He thinks it is important that every person in the world should be treated equally. He believes that everyone should have equal opportunities in life (i.e., item C in ESS questionnaire named ipeqopt).

(2) It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them (i.e., item H in ESS questionnaire named ipudrst).

(3) He strongly believes that people should care for nature. Looking after the environment is important to him (i.e., item S in ESS questionnaire named impenv).

Benevolence[**]

(1) It is essential for him to help the people around him. He wants to take care for their well-being (i.e., item L in ESS questionnaire named iphlppl).

(2) It is important for him to be loyal to his friends. He wants to devote himself to people close to him (i.e., item R in ESS questionnaire named iplylfr).

The ESS uses the Schwartz's Portrait Value Questionnaire (Schwartz, 2004; Schwartz et al., 2001) with the unipolar 6-point Likert scale (i.e., from 1 = very like me to 6 = not like me at all) to measure the aforementioned items.

The structural equation modeling (SEM) analyses will be conducted using LISREL v.9.30 (Jöreskog & Sörbom, 2017).

## 3. RESULTS

The general SEM model's fit was assessed using the classical goodness-of-fit indexes: the maximum likelihood ratio chi-square test, the goodness-of-fit index (GFI), and standardized root-mean-square residual (SRMR) as absolute goodness-of-fit indexes; the root-mean-square error of approximation (RMSEA) as parsimonious fit index; and the comparative fit index (CFI) and the non-normed fit index (NNFI) as incremental fit indices. Most of the SEM scientific community (Fan et al., 2016; Hu & Bentler, 1999; Kline, 2011; Schermelleh-Engel et al., 2003) suggests cutoff values of the aforementioned fit indexes: (a) low and not significant chi-square values are symptoms of good fit even though they are often found significant owing to the well-known limitations of this index, which is sensible to sample size. However, the chi-square magnitude is always reported as the first indication of discrepancy between the data and the hypothesized model; (b) values of RMSEA equal to or less than 0.05 are a good fit, in the range between 0.05 and 0.08 marginal, and greater than 0.10 is a poor fit; (c) GFI is similar to the coefficient of determination used in linear regression but applied to the entire model, and it reveals the amount of variance and covariance explained by the model (Bollen, 1989); (d)

[**] For simplicity's sake only two domains of the Schwartz' taxonomy have been selected, but the analyses can be expanded to the complete taxonomy or considering other domains of interest. It does not jeopardize the imaginary latent process.

SRMR values below 0.09 are considered good data-model fit; and (e) values greater than 0.90 for CFI and NNFI are considered adequate for a good model fit, although values approaching and over 0.95 are preferred.

Tables 2 and 3 present the CFA results of the Universalism and Benevolence latent Schwartz domains tested for the imaginary process with the maximum likelihood (ML) method of estimation[††] and the bootstrapping analysis[‡‡] on the constrained covariation matrix for testing the estimation stability caused by the sampling fluctuation. The first columns of both tables show the CFA solutions with no restrictions unless the first item is fixed to 1 to measure the respective latent as scaling indicators to identify the model (Bollen, 1989). This initial model with an effective sample size of 2546 respondents and 4 degrees of freedom performed fairly well regarding factor loadings (all over .5 and statistically significant from 0) and fit indices (i.e., chi square = 67.10 ($p$ < .000); GFI = .99; RMSEA = .079 with 90% confidence interval [.063–.096]; CFI = .98; NNFI = .96; SRMR = .02). This CFA model is the one to be tested for an imaginary process. Starting from Table 2, the Universalism is investigated as an imaginary value domain first with constraining its latent variance to –1[§§]. This process was repeated by selecting each item as scaling indicator alternatively to test for each item deficiency[***]. Therefore, the item coded impenv (i.e., He strongly believes that people should care for nature. Looking after the environment is important to him.) seems to be the only one found to be more resilient (i.e., factor loadings are greater) than the other two in the presence of a potential deficiency of the Universalism domain in Italy concerning the ESS sample. Practically, this means that for these citizens, a deficiency in Universalism will more likely affect their relationships with other people than their concern for preserving the environment. Passing to the Benevolence domain from Table 3 is straightforward, indicating that the most resilient item at a potential deficiency seems is the iplylfr (i.e., It is important to him to be loyal to his friends. He wants to devote himself to people close to him.) even though the bootstrapping solution did not confirm owing to the sampling fluctuation. However, these two items require further attention and investigation because they seem to preserve their own purposes, whereas Universalism and Benevolence concepts are more and more tenuous. Attention may regard, for instance, the context from which the items were surveyed, the research questions of the study, the characteristics of the sample, and so forth.

[††] Robust Maximum Likelihood (RML) and Robust Diagonally Weighted Least Square (DWLS) methods of estimation have been performed for considering also the potential ordinal nature of the variables (Finney & DiStefano, 2013), but here I just reported the ML solutions because they did not substantially differ from the other two strategies. All the RML and DWLS solutions are not reported, but they can be requested to the author.

[‡‡] The number of bootstrap samples was of 1000 (Hair et al., 2018) with 100% resampling of the raw data.

[§§]The SIMPLIS syntax, a program language that works under LISREL (Jöreskog & Sörbom, 2017) ambient, has been reported in the Appendix.

[***]Goodness-of-fit indices of the constrained model obviously got worse, even for bootstrapping, than the unconstrained solution because imposing a latent variance to be –1 computationally sounds improper (the worst example of fit indices found: chi square = 2462.02 ($p$ < .000); GFI = .73; RMSEA = .439 with 90% confidence interval (.425–.454); CFI = .40; SRMR = .33; the reader can easily run the CFAs reported in Table 2 with the SIMPLIS syntax provided in the Appendix). All that was expected and the goodness-of-fit indices here are not very informative because the purpose was not to find a good adaptation of original data matrix to the model-implied matrix but to look at the modifications of the indicators' common variances (i.e., factor loadings) while imposing an imaginary constraint.

**Table 2.** *Unstandardized (Std) factor loadings, latent variances, and covariances for Universalism as imaginary latent (*not significant at the 95% confidence level). Fixed values are indicated in bold. Bootstrapping results are indicated in italics.*

| UNIVERSALISM | | | | |
|---|---|---|---|---|
| | | Latent Variance | | |
| | 0.41 (1.00) | **-1.00** | **-1.00** | **-1.00** |
| ipeqopt | **1.00** (.64) | **1.00** (.96)<br>*1.00 (1.02)* | -.07 (-.07)<br>*-.12 (-.12)* | -.01* (-.01)<br>*-.02 (-.02)* |
| ipudrst | .97 (.66) | -.05 (-.05)<br>*-.09 (-.10)* | **1.00** (1.03)<br>*1.00 (1.10)* | .01* (.01)<br>*.01*(.01)* |
| impenv | 1.06 (.73) | .09 (.09)<br>*.08 (.08)* | .08 (.09)<br>*.08 (.08)* | **1.00** (1.07)<br>*1.00 (1.08)* |
| BENEVOLENCE | | | | |
| | | Latent Variance | | |
| | .43 (1.00) | .43 (1.00)<br>*.43 (1.00)* | .46 (1.00)<br>*.43 (1.00)* | .38 (1.00)<br>*.36 (1.00)* |
| iphlppl | **1.00** (.70) | **1.00** (.71)<br>*1.00 (.71)* | **1.00** (.73)<br>*1.00 (.71)* | **1.00** (.67)<br>*1.00 (.66)* |
| iplylfr | 1.00 (.73) | .98 (.72)<br>*.98 (.73)* | .92 (.70)<br>*.98 (.73)* | 1.12 (.77)<br>*1.15 (.79)* |
| UNIVERSALISM-BENEVOLENCE | | | | |
| | | Latent Covariance | | |
| | .42 (1.00) | .45 (.68)<br>*.38 (.58)* | .43 (.64)<br>*.35 (.54)* | .43 (.70)<br>*.41 (.68)* |

**Table 3.** *Unstandardized (Std) factor loadings, latent variances, and covariances for Benevolence as imaginary latent. (*not significant at the 95% confidence level). Fixed values are indicated in bold. Bootstrapping results are indicated in italics*

| BENEVOLENCE | | | |
|---|---|---|---|
| | | Latent Variance | |
| | 0.41 (1.00) | **-1.00** | **-1.00** |
| iphlppl | **1.00** (.70) | **1.00** (.1.06)<br>*1.00 (1.09)* | .01* (.01)<br>*-.02 (-.02)* |
| iplylfr | 1.00 (.73) | .02 (.02)<br>*-.00* (-.00)* | **1.00** (1.11)<br>*1.00 (1.16)* |
| UNIVERSALISM | | | |
| | | Latent Variance | |
| | .43 (1.00) | .44 (1.00)<br>*.43 (1.00)* | .40 (1.00)<br>*.39 (1.00)* |
| ipeqopt | **1.00** (.64) | **1.00** (.66)<br>*1.00 (.67)* | **1.00** (.64)<br>*1.00 (.64)* |
| ipudrst | .97 (.66) | .98 (.68)<br>*.94 (.68)* | .95 (.64)<br>*.94 (.64)* |
| impenv | 1.06 (.73) | .97 (.69)<br>*.98 (.69)* | 1.10 (.75)<br>*1.14 (.76)* |
| BENEVOLENCE - UNIVERSALISM | | | |
| | | Latent Covariance | |
| | .42 (1.00) | .45 (.69)<br>*.41 (.63)* | .43 (.67)<br>*.39 (.63)* |

## 4. DISCUSSION and CONCLUSION

Recalling Rindskopf (1984, p.38), the imaginary LVs should be variables useful to implement specific constraints in measurement models. Above and beyond this initial definition and based on the empirical test provided in this manuscript, it can be reasonable to propose that imaginary LVs are variables useful for testing a latent deficiency within a specific context of the application. Explicitly, the imaginary LVs while postulating variances equal to −1 reflect this negative effect within their observed indicators that turn into complex numbers. Consequently, the measurement equations of confirmatory factor models with imaginary LVs turn into measurement equations with complex numbers. However, on one hand, solving these new complex equations with the usual SEM techniques yields expected unacceptable fit indices; on the contrary, it still provides significant structural parameters and thus potential indications on which indicator, loading the imaginary latent, is less (or more) affected by this latent deficiency. That is to say, because a negative latent variance is a variability that is absent in a latent concept, this sort of latent lacking will be reflected in the indicators, and thus, it can sensibly give signals on what would happen if that latent concept is flawed: which latent aspect (measured by each indicator) will be more affected by, and which is more resilient to, this potential deficiency. These potential indications need to be more investigated and/or validated by other SEM-based strategies (like measurement invariance across groups for instance), but I strongly suggest that it is something not to be ignored. This empirical test can also add further potential information on the selection of scaling indicators while a deficiency scenario in the LVs is hypothesized and therefore contributes to expanding the list of criteria for this selection (Bollen et al., 2022).

Furthermore, and perhaps most importantly, this empirical test of the imaginary latent interrogation opens new possibilities regarding the promising usefulness of the complex numbers in measurement models with latent variables that, to my knowledge, are still unexplored and so are the subsequent estimation methods of these types of SEM models. While using the well-known methods of estimation (e.g., ML, RML, DWLS), a researcher obtains bad fit indices because you are running models with offending constraints like fixing latent variances to –1. Consequently, new methods, possibly even completely different from the usual ones, that include the math process of imaginary and complex numbers in the estimation process are eagerly necessary, although it goes beyond the purpose of this work that remains essentially pioneering. However, the two-factor model tested in this initial experiment yielded promising results that warrant further investigation, particularly involving multifactor structures with additional reflective items to be tested across different respondent groups.

Finally, the evident limitations of this approach need to be considered. The first was just partially mentioned above and regards the methodological way how to model an imaginary latent. In this experiment, the LISREL computational system was pragmatically forced to converge to a solution by fixing the variance of a latent variable to be equal to –1. Other statistical software like M-Plus (Muthen & Muthen, 1998-2017) and lavaan (Rosseel, 2012) under R (R core Team, 2021) can be tried, but I am more than certain that other methods of estimation are needed. A second limitation is that only reflective indicators have been tested for potential deficiency in a latent variable. However, what happens when formative causal indicators are included? They are typical predictors of a latent variance such as the multiple indicators multiple causes (MIMIC) model (Jöreskog & Goldberger, 1975) (Bollen & Diamantopoulos, 2015). Whenever an imaginary interrogation is requested for latent variable models with formative indicators, it would mean that they predict a negative latent variance by estimating possible causes behind the deficiency found in the relative reflective indicators. This sounds like another extremely challenging perspective to be explored in the future.

## Acknowledgments

The author may want to thank both reviewers very much for their excellent remarks. Special thanks to the reviewer #1 for his/her comments and suggestions that substantially improved this manuscript.

## Declaration of Conflicting Interests and Ethics

The author declares 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.

## Authorship Contribution Statement

**Marco Vassallo**: Conceived the presented idea, conducted the formal analyses, and wrote the article.

## Orcid

Marco Vassallo https://orcid.org/0000-0001-7016-6549

## REFERENCES

Bentler, P.M., & Lee, S.Y. (1983). Covariance structures under polynomial constraints: Applications to correlation and alpha-type structural models. *Journal of Educational Statistics*, *8*, 207–222. https://doi.org/10.2307/1164760

Bollen, K.A. (1989). Structural equations with latent variables. New York NY, USA: Wiley Press.

Bollen, K.A., & Diamantopoulos, A. (2015). In defense of causal–formative indicators: A minority report. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000056

Bollen, K.A., Lilly, A.G., & Luo, L. (2022). Selecting scaling ındicators in structural equation models (SEMs). *Psychological Methods*, Advance online publication. http://dx.doi.org/10.1037/met0000530

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York NY, USA: The Guilford Press.

ESS Round 10: European Social Survey. (2022). ESS-10 2020 Documentation Report. Edition 1.0. Bergen, European Social Survey Data Archive, Sikt - Norwegian Agency for Shared Services in Education and Research for ESS ERIC, Norway. https://doi:10.21338/NSD-ESS10-2020

ESS Round 10: European Social Survey Round 10 Data. (2020). Data file edition 1.2. Sikt - Norwegian Agency for Shared Services in Education and Research, Norway – Data Archive and distributor of ESS data for ESS ERIC, Norway. https://doi:10.21338/NSD-ESS10-2020

Fabricar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S.R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(19). https://doi.org/10.1186/s13717-016-0063-3

Finney, S.J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (2nd ed., pp. 439-492). Greenwich, CT, Information Age Publishing.

Hair, J.F., Sarstedt, M., Ringle, C.M., & Gudergan, S.P. (2018). *Advanced Issues in Partial Least Squares Structural Equation Modeling*. Thousand Oaks, CA, USA: Sage.

Hancock, G.R., Stapleton, L.M., & Arnold-Berkovits, I. (2009). The tenuousness of invariance tests within multi-sample covariance and mean structure models. In T. Teo & M.S. Khine

(Eds.), *Structural Equation Modeling in Educational Research: Concepts and Applications* (pp. 137-174). Rotterdam, Netherlands: Sense Publishers.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit ındexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Jöreskog, K.G. (1966). Testing a simple structure hypothesis in factor analysis. *Psychometrika*, *31*, 165-178. https://doi.org/10.1007/BF02289505

Jöreskog, K.G., & Goldberger, A.S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of American Statistical Association*, *70*, 631-639. https://doi.org/10.2307/2285946

Jöreskog, K.G., & Sörbom, D. (2017). *LISREL 9.30 for Windows*. [Computer software manual]. Scientific Software Skokie, IL, USA: International Inc.

Kline, B.R. (2011). *Principles and Practice of Structural Equation Modeling, 3rd ed*. New York, NY, USA: The Guilford Press.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading MA, USA: Addison-Wesley press.

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition*. [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

R Core Team. (2014). R*: A language and environment for statistical computing*. [Computer software manual]. http://www.R-project.org/

Rindskopf, D. (1984). Using phantom and imaginary latent variables to parameterize constraints in linear structural models. *Psychometrika*, *49*, 37-47. https://doi.org/10.1007/BF02294204

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. http://www.jstatsoft.org/v48/i02/

Schermelleh-Engel, K., Moosbruger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive-of-fit measures. *Methods of Psychological Research*, *8*, 23–74. https://doi.org/10.23668/psycharchives.12784

Schwartz, S.H. (1992). Universals in the content and structure of values: Theoretical advance and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, *25*, 1-65.

Schwartz, S.H. (2004). Basic human values: Their content and structure across countries. In A. Tamayo & J. Porto (Eds.). *Valores e Trabalho (Values and Work)*. Brasilia, Brasile: Editora Universidade de Brasilia.

Schwartz, S.H., Melech, G., Lehman, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross Cultural Psychology*, *32*, 519-542. https://doi.org/10.1177/0022022101032005001

Thurstone, L.L. (1947). *Multiple-factor analysis*. Chicago, USA: University of Chicago Press.

Weisstein, E.W. (2023). *Argand Diagram*. From MathWorld [A Wolfram Web Resource]. *https://mathworld.wolfram.com/ArgandDiagram.html*

## APPENDIX

SIMPLIS syntax to run the maximum likelihood analysis (no bootstrapping) of the path model in Table 2 and 3 within the main text; "!" stands for comments. The reader can alternatively set the variance of U (Universalism) or B (Benevolence) to -1 to check the U and B items, respectively):

```
Observed variables ipeqopt ipudrst impenv iphlppl iplylfr
Covariance Matrix
0.997
0.444 0.896
0.427 0.401 0.870
0.427 0.431 0.431 0.860
0.402 0.384 0.485 0.428 0.804
Latent variables U B
Sample Size = 2546
Relationships
ipeqopt=1*U
ipudrst=U
impenv=U
iphlppl=1*B
iplylfr=B
Set Variance of U to -1 ! Set Variance of B to -1
Path Diagram
Print Residuals
Admissibility check = off
End of Problem
```

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 733–757
https://doi.org/10.21449/ijate.1327238
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Redefining the impact of professional development in education with ProDES (Professional Development Evaluation Scale)

**Mustafa Özgenel[1*], Martin Brown[2], Joe O'Hara[2], Metin Özkan[3]**

[1]Istanbul Sabahattin Zaim University, Faculty of Education, Istanbul, Türkiye
[2]Dublin City University, Institute of Education, Dublin, Ireland
[3]Gaziantep University, Faculty of Education, Gaziantep, Türkiye

**ARTICLE HISTORY**

*Received: July 13, 2023*

*Accepted: Aug. 26, 2024*

**Keywords:**

Professional development,
Assessment of professional development,
Scale development,
Professional learning.

**Abstract:** This study introduces the Professional Development Evaluation Scale (ProDES), a tool that has been developed to evaluate the impact of professional development as it relates to participants' Learning and Use of New Knowledge and Skills, Organisational Support, Student Learning Outcomes, and reactions. Grounded in Guskey's (2000) framework for evaluating Professional Development, ProDES was developed with data from five study groups in Turkey and underwent refinement across four factors. Exploratory and Confirmatory Factor Analyses confirmed the scale's structure, accounting for 62.72% of the total variance, with robust fit indices. Within this, ProDES demonstrated high internal consistency and test-retest reliability, with significant correlations validating its effectiveness. The scale's high internal consistency and test-retest reliability ensure that it can be used to make evidence-informed decisions that can foster more effective and supportive professional development activities. As a result, by identifying which professional development initiatives lead to improvements, those associated with professional development can use resources more efficiently, leading to enhanced school and system-wide improvements. Moreover, the use of ProDES can also help schools and education systems track progress over time, making ProDES an invaluable tool for continuous improvement and strategic planning.

## 1. INTRODUCTION

Education policymakers, researchers, and practitioners recognize the crucial role of professional development for school administrators and teachers (educational professionals responsible for the management and leadership of schools such as school principals, assistant principals, Heads of Departments, and inspectors) (Bredeson, 2000). Indeed, teachers and administrators require ongoing professional development to sustain their current professional skills, knowledge, and competencies and require new skills due to the changing roles and responsibilities they face as well as shortcomings in their pre-service training (Spillane et al., 2009). In other words, professional development serves as a strategy and policy tool for school

*CONTACT: Mustafa ÖZGENEL ✉ mustafa.ozgenel@izu.edu.tr İstanbul Sabahattin Zaim University, Faculty of Education, Department of Educational Sciences, İstanbul, Türkiye



**e-ISSN: 2148-7456**

improvement, with the assumption that practitioners need to acquire new knowledge and skills (Guskey, 2002a).

It is considered a fundamental component of successful educational reform and school improvement that should ultimately lead to, for example, changes in teachers' classroom practice, attitudes, beliefs, and student learning outcomes (Little, 1993). Professional development can also be viewed as a transformative process that encourages the attainment of high-value goals (Assor & Oplatka, 2003) as well as a lifelong as opposed to a once-off event for educators to meet the ever-changing needs of students (Diaz-Maggioli, 2004). In the case of education, it is conceptualized as a form of adult learning that supports the learning of administrators and teachers (Zepeda, 2011).

Desimone (2011) defines professional development as a complex series of interconnected learning opportunities, while Guskey (2000) describes it as a process and activities designed to enhance educators' professional knowledge, skills, and attitudes to promote the advancement of their students. Fullan (1994) has also highlighted the importance of investing in teachers' professional development for the implementation of planned change strategies. Indeed, there is a long-held view that the professional development of teachers and administrators should be at the heart of all plans and policies for school improvement (Adey, 2004; Barth, 1986; Blandford, 2004; Bredeson, 2000; Hallinger, 2003). In other words, for school development and quality education, it is an essential process for administrators and teachers to improve, change, and adapt their attitudes and behaviors (Easton, 2008), as well as enhance their knowledge, skills, and competencies. Although the potential contribution of professional development to school success has been well documented, and several studies have explored its causal effects (e.g., Garet et al., 2001; Wayne et al., 2008), research on its impact often tends to either theorize about it or rely on Quod Erat Demonstrandum (QED) findings. These studies frequently highlight the success of initiatives implemented by the researchers, often featuring overly positive accounts from participants involved. To concur with Bredeson (2000), it is not possible to demonstrate the impact of professional development without robust empirical evidence, which we strongly suggest is lacking in the professional development literature. In this regard, ethically sound "evaluation" of professional development activities or programs provided to administrators and teachers is significant and should be an integral part of professional development activities (Blandford, 2012).

To genuinely evaluate the impact of professional development, five critical pieces of information are needed: *(i)* participants' reactions to the professional development experience, *(ii)* knowledge and skills acquired by participants, *(iii)* school support for professional development, *(iv)* participants' use of newly acquired knowledge and skills in their professional practices, and *(v)* evidence of how professional development activities impact and benefit students (Guskey, 1999, 2000, 2002a). In this regard, high-quality data collection tools specifically designed to assess teachers' and administrators' perceptions or attitudes towards professional development activities or programs are required. Numerous data collection instruments have been developed or adapted to measure teachers' various aspects of professional development and in different contexts such as:

- self-efficacy (Yenen & Kılınç, 2021),
- attitudes towards professional development (Torff, Sessions, & Byrnes, 2005; Özer & Beycioğlu, 2010),
- perceptions of professional development (Mourão et al., 2014; Soine & Lumpe, 2014),
- continuous professional development of social and health care educators (Koskimäki et al., 2021),
- professional development needs (Shabani et al., 2018),
- factors influencing professional development processes (Saberi & Sahragard, 2019),

- participation in professional learning (Liu, Hallinger, & Feng, 2016; Gümüş, Apaydın, & Bellibaş, 2018),
- teachers' motivation for web-based professional development (Çakır & Horzum, 2014; Kao, Wu, & Tsai, 2011),
- professional development activities (Dijkstra, 2009; Eroğlu & Özbek, 2018; Eroğlu & Özbek, 2020; Kwakman,1999) and
- pre-service teachers' professional development (Zhu, 2015).

However, these data collection tools do not specifically focus on assessing teachers' and administrators' professional development. Additionally, there is a lack of data collection tools rooted in Guskey's (1999, 2000, 2002b) theoretical framework that directly address the evaluation of professional development. Therefore, the professional development scale developed in this research, referred to as the Professional Development Evaluation Scale (ProDES), has been specifically developed to address a gap in professional development research. More specifically, the scale that can be used by researchers, schools and professional development service providers can be used to evaluate:

- participants reactions to professional development activities/initiatives.
- participants acquired knowledge, skills and competencies gained from these activities.
- participants use of what they have learned/gained in their professional practices.
- the impact of the professional development on student learning outcomes.
- perceptions of organizational/administrative support.

## 2. METHOD

### 2.1. Research Design

This study was designed and conducted according to the survey design to develop a measurement tool to evaluate the professional development of teachers and administrators.

### 2.2. Study Groups and Data Collection

After obtaining ethical approval from the Istanbul Sabahattin Zaim University Research and Publication Ethics Committee (Decision No:2023-2023/04), data were collected through face-to-face interviews and Google Forms. For the purpose of developing the ProDES scale, this study involved voluntary participation of administrators and teachers. Data was collected using surveys from five study groups and included trial testing, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), test-retest reliability, and criterion validity analyses. Using convenience sampling, teachers and administrators working in public and state schools during the 2022-2023 academic year were selected for the development of ProDES. The choice of sampling strategy used was based on the requirement of obtaining the desired sample size for responding to a measurement tool (Robson, 2017).

The determination of the sample size for EFA and CFA is a topic of much debate in the literature. It is stated that a sample size of 5 to 10 times the number of items may be sufficient for factor analysis in scale development studies (Hair et al., 1998; Ho, 2006; MacCallum et al., 1999). Tabacknick and Fidell (2001), on the other hand, suggest that a sample size of at least 300 is appropriate. In this study, 586 and 478 participants were included in the EFA and CFA, respectively. Therefore, in line with the literature, the sample size used for the development of ProDES was sufficient for EFA and CFA. The information regarding the study groups used in the analysis is presented in Table 1.

When examining Table 1, the majority of participants in the study groups were female teachers working as teachers at the elementary school level. Additionally, the average age of participants in the pilot phase of the study was ±42.90 (*sd*=9.011), the average teaching experience was ±19.26 years (*sd*=8.594), and the average administrative experience was ±1.69 years (*sd*=4.241). For EFA, the average age of participants was ±37.17 (*sd*=8.497), the average

teaching experience was ±13.20 years (*sd*=8.398), and the average administrative experience was ±6.46 years (*sd*=5.581). For CFA, the average age of participants was ±39.59 (*sd*=8.412), the average teaching experience was ±15.41 years (*sd*=8.412), and the average administrative experience was ±755 years (*sd*=6.418). In terms of criterion validity, the average age of the participants was ±42.19 (*sd*=9.044), the average teaching experience was ±18.42 years (*sd*=8.855), and the average administrative experience was ±1.16 years (*sd*=4.689). For the test-retest application, the average age of the participants was ±43.78 (*sd*=9.333), the average teaching experience was ±19.26 years (*sd*=8.731), and the average administrative experience was ±1.45 years (*sd*=4.437). In summary, therefore, it can be observed that the study groups involved in the development of ProDES exhibit a heterogeneous structure in terms of age, experience, school type, and level of work.

**Table 1.** *Distribution of the study group according to demographic variables.*

| | | Trial | | EFA | | CFA | | Criterion Validity | | Test-retest | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Groups | | *f* | % | *f* | % | *f* | % | *f* | % | *f* | % |
| Gender | Female | 86 | 57.7 | 411 | 7.1 | 248 | 51.9 | 99 | 58.2 | 28 | 59.6 |
| | Male | 63 | 42.3 | 175 | 29.9 | 230 | 48.1 | 71 | 41.8 | 19 | 4.4 |
| School Type | Preschool | 12 | 8.1 | 45 | 7.7 | 44 | 9.2 | 22 | 12.9 | 2 | 4.3 |
| | Elementary | 52 | 34.9 | 238 | 4.6 | 208 | 43.5 | 64 | 37.6 | 45 | 95.7 |
| | Middle | 46 | 30.9 | 177 | 3.2 | 140 | 29.3 | 34 | 2.0 | - | - |
| | High | 39 | 26.2 | 116 | 19.8 | 82 | 17.2 | 50 | 29.4 | - | - |
| | Other | - | - | 10 | 1.7 | 4 | .8 | - | - | - | - |
| Position | Teacher | 133 | 89.3 | 515 | 87.9 | 413 | 86.4 | 149 | 87.6 | 43 | 91.5 |
| | Deputy P. | 14 | 9.4 | 52 | 8.9 | 33 | 6.9 | 18 | 1.6 | 3 | 6.4 |
| | Principal | 2 | 1.3 | 19 | 3.2 | 32 | 6.7 | 3 | 1.8 | 1 | 2.1 |
| Total | | 149 | 100.0 | 586 | 100.0 | 478 | 100.0 | 170 | 100.0 | 47 | 100.0 |

## 2.3. Scale Development Process

In the process of scale development, the steps recommended by Hinkin (1998) and Hinkin et al. (1997) were followed, including *(i)* item writing, *(ii)* content validity, *(iii)* determination and implementation of the sample size, (*iv)* exploratory and confirmatory factor analysis, *(v)* internal consistency/reliability, and *(vi)* criterion validity determination.

As part of the scale development process, a systematic literature review was conducted to examine the professional development of teachers and administrators (e.g., Campbell et al., 2004; Cohen, 2004; Guskey, 2003a, 2003b; Guskey & Yoon, 2009; Kirkpatrick & Kirkpatrick, 2006). Existing measurement tools developed or adapted into Turkish in this field were examined (Çakır & Horzum, 2014; Eroğlu & Özbek, 2018, 2020; Eroğlu, 2019; Gümüş et al., 2018; Koskimäki et al., 2021; Mourão et al., 2014; Saberi & Sahragard, 2019; Shabani et al., 2018; Torff et al., 2005; Yenen & Kılınç, 2021; Zhu, 2015).

A pool of 72 items was created and based on Guskey's (1999, 2000, 2002b) 5-level professional development evaluation model. During the item writing process, attention was given to ensuring that the items assessed the activities/programs that administrators and teachers engaged in for their professional development, focused on evaluating a single behaviour or action, avoided misinterpretation, and used expressions that the target audience could understand in terms of language and meaning. After checking the items, eight items were removed, and a draft form with 64 items was emailed to five experts in measurement evaluation, seven experts in educational administration, one expert in early childhood education, one expert in linguistics, and one expert in program development. These experts were provided with explanations about the research purpose and scale and were asked to evaluate each item. Based on feedback from these experts, the content validity ratio (CVR) and content validity index (CVI) suggested by Lawshe (1975) was calculated for each item. A trial sample was conducted

with 149 participants using the 43-item version. The final version of the scale was determined based on the feedback received from the target group during the trial implementation.

### 2.4. Data Analysis

Data obtained from 586 participants was used for EFA. To determine whether the data was suitable for factor analysis, assumptions such as outliers, missing values, normality, multicollinearity, and sufficient sample size were examined.

To detect outliers, z-scores were calculated for all individuals, and it was observed that they fell within the range of -2.58 to +1.90. No data points were outside the ±3 range (Tabachnick & Fidell, 2001). P-P plot, skewness, and kurtosis coefficients were also examined to check the assumption of normality. The item scores in the dataset had skewness and kurtosis values within the range of ±1.00. According to Çokluk et al. (2012), when the skewness and kurtosis values are within the ±1 range, the data are considered to follow a normal distribution.

Collinearity issues were examined through Pearson Product-Moment Correlation between the items, and a comparison of the lower and upper 27% groups was conducted to assess the discriminant validity of the items. Each item had a *t*-value greater than ±1.96, item-total correlation values ranged from $r$=.353 to $r$=.776, and there were no multicollinearity issues ($p$<.01). Bartlett's test of sphericity and Kaiser-Meyer-Olkin (KMO) tests were also used to test the suitability of the sample size and data for factor analysis.

Bartlett's test of sphericity was significant, and the KMO value was close to 1, indicating that the data were suitable for factor analysis. EFA began with Principal Component Analysis, followed by the Varimax Rotation Technique. In the analyses, an item loading estimation point of 0.50 was used, and items with loading values below 0.50 and items with cross-loadings on multiple factors were sequentially removed, of which the analyses were repeated after each item was removed. Factor loadings are ideally expected to exceed 0.40, particularly within multidimensional frameworks (Howard, 2016). Given that a substantial factor loading indicates a heightened association between an item and its corresponding factor (Kılıç, 2022), a factor loading threshold of 0.50 was employed in the present study.

CFA was conducted using data collected from 478 participants to confirm the 4-factor structure consisting of the 29 items identified in the EFA. Assumptions were tested to assess the suitability of the CFA data. There were no missing values in the dataset, the z-scores of the data ranged from -3 to +3, the skewness and kurtosis values of the item scores were within ±1.00, and the Pearson Product-Moment Correlations between the items were less than 0.80. Therefore, the dataset met the assumptions of no outliers, normality, and multicollinearity. The maximum likelihood (ML) method was used for parameter estimation of the CFA model. The utilisation of the Maximum Likelihood estimation method was prioritized in this study due to the normal distribution of the data and the attainment of a sizable sample. Maximum Likelihood is favoured for yielding more dependable parameter estimates under conditions where the assumptions are satisfied, and a substantial sample size is achievable, as stated by Helm Castro-Schilo and Oravecz (2017).

Item-total and item-rest correlations were also examined to determine the discriminant validity between items that measured the intended constructs and those that did not. Additionally, *t*-tests comparing the lower and upper 27% groups were conducted. Furthermore, to provide evidence for criterion validity, correlation values between the Professional Development Attitude Scale and the scale developed in this study were calculated for a study group consisting of 170 participants, with a three-week interval between measurements.

The scale developed by Torff et al. (2005) to measure teachers' attitudes towards professional development was adapted into Turkish by Özer and Beycioğlu (2010). The original scale consisted of nine items and a single dimension; however, in the Turkish adaptation, three items were removed from the scale. The items in the 5-point Likert scale are rated on a range from

"*Strongly Disagree*=1" to "*Strongly Agree*=5." The second item of the scale, "I consider the money spent on professional development programs for teachers to be wasted," was reverse scored. Cronbach's alpha reliability coefficient for this study's scale was 0.778. Additionally, the goodness-of-fit indices for confirmatory factor analysis were examined [$\chi^2/df$= 19.603/8= 2.450; RMR= .093; SRMR= .038; GFI= .965; AGFI= .907; IFI= .965; CFI= .964; RMSEA= .093], and it was observed that the values were within acceptable limits.

Reliability requires that a measurement or measurement tool consistently reflects the construct it measures (Field, 2009). For this reason, reliability coefficients with different theoretical and statistical procedures were calculated (George & Mallery, 2009), and Cronbach's alpha, McDonald's Omega, Split-half, Equivalent forms, Guttman, and Sperman-Brown reliability coefficients were presented as evidence. Finally, Jamovi, IBM SPSS, and IBM AMOS software packages were used for data analysis. The significance level for statistical analysis was set at .05.

## 3. FINDINGS

### 3.1. Content Validity

Table 2 presents the calculated Content Validity Ratio (CVR) values for each item and the overall Content Validity Criterion (CVI) values obtained for the entire scale.

**Table 2.** *Lawshe's analysis results.*

| Items | CVR | Items | CVR | Items | CVR | Items | CVR |
|---|---|---|---|---|---|---|---|
| 1 | 0.867 | 12 | 0.733 | 23 | 0.600 | 34 | 1.000 |
| 2 | 0.600 | 13 | 0.867 | 24 | 0.867 | 35 | 0.733 |
| 3 | 1.000 | 14 | 0.733 | 25 | 0.867 | 36 | 0.733 |
| 4 | 0.867 | 15 | 1.000 | 26 | 0.867 | 37 | 0.867 |
| 5 | 0.857 | 16 | 1.000 | 27 | 0.867 | 38 | 1.000 |
| 6 | 1.000 | 17 | 0.867 | 28 | 1.000 | 39 | 1.000 |
| 7 | 1.000 | 18 | 0.867 | 29 | 0.857 | 40 | 0.867 |
| 8 | 1.000 | 19 | 0.733 | 30 | 0.867 | 41 | 0.867 |
| 9 | 1.000 | 20 | 0.733 | 31 | 0.867 | 42 | 1.000 |
| 10 | 0.867 | 21 | 0.867 | 32 | 1.000 | 43 | 0.733 |
| 11 | 0.867 | 22 | 0.867 | 33 | 1.000 | | |
| Content Validity Index (CVI)=0.860, Content Validity Ratio (CVR-N=15): 0.49 | | | | | | | |

According to the comparison based on Lawshe's (1975) recommended content validity ratio, in this study, the critical value for CVR was determined as 0.49 at a significance level of $p$=0.05. Consequently, 21 items were removed from the draft form based on this critical value. The CVI for the remaining 43 items was 0.86. In this sense, it can be concluded that the scale provides content validity.

### 3.2. Item Analysis

Prior to EFA, the total score for each participant was obtained. The upper 27% of the entire group, consisting of 158 participants, was selected as the high-score group, whereas the lower 27% of the entire group, consisting of 158 participants, was selected as the low-score group. Subsequently, t-test was performed between the two groups, and the t-test results are presented in Table 3.

When examining Table 3, it can be observed that the $t$-values of the items are significant ($p$<.01), and the $t$-values are greater than 1.96. Conducting item analysis and selecting the items that contribute the most to the scale enhances its validity (Erkuş, 2014; Özgüven, 2015). In this regard, it can be inferred that the items in the 43-item draft form were suitable for factor analysis.

**Table 3.** *The t-value (Item discrimination index) of the items in the* ProDES.

| Items | *t*-value | Items | *t*-value | Items | *t*-value | Items | *t*-value |
|---|---|---|---|---|---|---|---|
| M1 | -16.987 | M12 | -20.536 | M23 | -24.416 | M34 | -19.133 |
| M2 | -18.002 | M13 | -19.525 | M24 | -21.849 | M35 | -19.505 |
| M3 | -18.973 | M14 | -20.690 | M25 | -20.310 | M36 | -18.754 |
| M4 | -14.657 | M15 | -15.246 | M26 | -20.127 | M37 | -17.538 |
| M5 | -8.013 | M16 | -15.159 | M27 | -21.036 | M38 | -18.063 |
| M6 | -9.403 | M17 | -14.733 | M28 | -19.964 | M39 | -16.987 |
| M7 | -16.874 | M18 | -21.134 | M29 | -20.399 | M40 | -19.552 |
| M8 | -17.659 | M19 | -14.462 | M30 | -17.367 | M41 | -15.682 |
| M9 | -19.477 | M20 | -14.118 | M31 | -18.941 | M42 | -15.534 |
| M10 | -19.096 | M21 | -19.942 | M32 | -21.697 | M43 | -6.114 |
| M11 | -20.270 | M22 | -23.387 | M33 | -20.558 | | |

### 3.3. Validity

Before conducting factor analysis, the sample size and suitability of the data for factor analysis were evaluated using measures of normality, the Kaiser-Meyer-Olkin Measure (KMO), Bartlett's Test of Sphericity, Nonadditivity, and Hotelling's $T^2$ Test.

**Table 4.** *The skewness and kurtosis coefficients of the datasets on which EFA and CFA were conducted.*

| | EFA | | CFA | |
|---|---|---|---|---|
| | Statistic | Std. Error | Statistic | Std. Error |
| Mean | 2.8964 | .02395 | 2.8462 | .02489 |
| Median | 2.9302 | | 2.8966 | |
| Variance | .336 | | .387 | |
| Std. Deviation | .57973 | | .62222 | |
| Minimum | 1.40 | | .31 | |
| Maximum | 4.00 | | 4.00 | |
| Range | 2.60 | | 3.69 | |
| Interquartile Range | .80 | | .76 | |
| Skewness | -.244 | .101 | -.528 | .098 |
| Kurtosis | -.491 | .202 | .711 | .195 |

According to Table 4, the collected data for EFA fell within the range of ±1 for the skewness and kurtosis coefficients. Following George and Mallery (2016), data is considered to exhibit a normal distribution when skewness and kurtosis coefficients fall between ±1. To assess the collectability of the draft scale, a non-additivity test was conducted. To evaluate the additivity of the draft scale, the additivity test (Table 5) and Hotelling T Test (Table 6) were performed to determine whether there was a significant difference between the item averages. While Tukey's test of additivity tests the linear dependence between variables; Hotelling's T-square tests whether the means of the variables are equal (George & Mallery, 2016).

**Table 5.** *The ANOVA Tukey test conducted for nonadditivity.*

| | Sum of Squares | *df* | Mean Square | *F* | *p* |
|---|---|---|---|---|---|
| Between groups | 8454.181 | 585 | 14.452 | 40.817 | .000 |
| Within groups | 897.071 | 42 | 21.359 | | |
| Nonadditivity | 58.218[a] | 1 | 58.218 | 111.757 | .000 |

When examining Table 5, it can be seen that the probability of nonadditivity is $p$=.000, indicating that the scale does not possess the property of additivity ($F$=111.757; $p$<.01). When

examining the variability between measurements, significant differences were observed, but it is understood that the scale does not possess the property of additivity ($F$=40.817; $p$<.01).

**Table 6.** *Hotelling's $T^2$ testi.*

| Hotelling's T-test square | *F* | *df*1 | *df*2 | *p* |
|---|---|---|---|---|
| 170.085 | 6.168 | 26 | 413 | .000 |

According to Table 6, it can also be observed that the item means are not equal to each other ($F$=6.168; $p$<.001). Since the item means show significant differences, this indicates that the items measuring different tendencies/attitudes/characteristics are perceived differently by a heterogeneous group, and the scale has more than one factor.

To perform factor analysis, it is recommended that the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy should be equal to or greater than 0.60, indicating an acceptable level of sampling adequacy. Additionally, Bartlett's test of sphericity should produce a statistically significant result, suggesting that the variables in the dataset were sufficiently correlated for factor analysis (Tabachnick & Fidell, 2007). A KMO value above 0.80 indicates that the data set obtained from the sample is "very good" (Tavşancıl, 2002), and a significant result of Bartlett's Test indicates that the data are derived from multivariate normal distribution (Otrar & Argın, 2015). The KMO and Bartlett's test values for the dataset in which EFA and CFA analyses were conducted are presented in Table 7.

**Tablo 7.** *KMO and Barlett's test.*

| | | EFA | CFA |
|---|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .963 | .963 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 16732.822 | 10952.581 |
| | *df* | 903 | 406 |
| | *p* | .000 | .000 |

When examining the suitability of the data for EFA and CFA in Table 7, it was determined that the KMO value is close to 1 and the result of Bartlett's Test is significant (EFA=$\chi^2$=16732.822, $df$=903, $p$<.001; CFA=$\chi^2$=10952.581, $df$=406, $p$<.001). These findings indicate that the sample size and data sets are sufficient for EFA and CFA (Hof, 2012; Tatlıdil, 2022). Validity allows us to obtain information about the property that the scale intends to measure (Thorndike & Thorndike-Christ, 2017). Therefore, Exploratory Factor Analysis (EFA) was conducted to identify the dimensions and number of factors, if any, related to the intended property of the scale (Brown & Moore, 2013). EFA begins with a principal component analysis. Eigenvalues are used to determine the factors (Tavşancıl, 2002). Eigenvalue indicates the amount of information obtained from a factor (DeVellis, 2014).

In factor analysis, factors with an eigenvalue of 1 or greater are included in the analysis (Büyüköztürk, 2012; Tavşancıl, 2002). In factor analysis, it is recommended to perform Varimax Rotation unless there are compelling reasons to determine the distribution of items across factors, as factor loading values of items affect the amount of explained variance, and it is desired to have high factor loading values for items (Büyüköztürk, 2002; Tabachnick & Fidell, 2007). In factor analysis, attention was paid to the factor loadings of items being at or above .50, items not loading on multiple factors, and a minimum difference of .10 between factor loading values for items loading on multiple factors (Çokluk et al., 2012; Tavşancıl, 2002). Following the principal component analysis, the Varimax Orthogonal Rotation technique was used, and no dimension restriction was applied in EFA to reveal the factor structure of the scale. Fourteen items were sequentially removed that had factor loadings below .50 and loaded on multiple factors (items 43, 11, 14, 13, 21, 4, 28, 18, 27, 12, 2, 6, 5, 29). Table 8 presents the factor eigenvalues and explained variance ratios obtained from EFA.

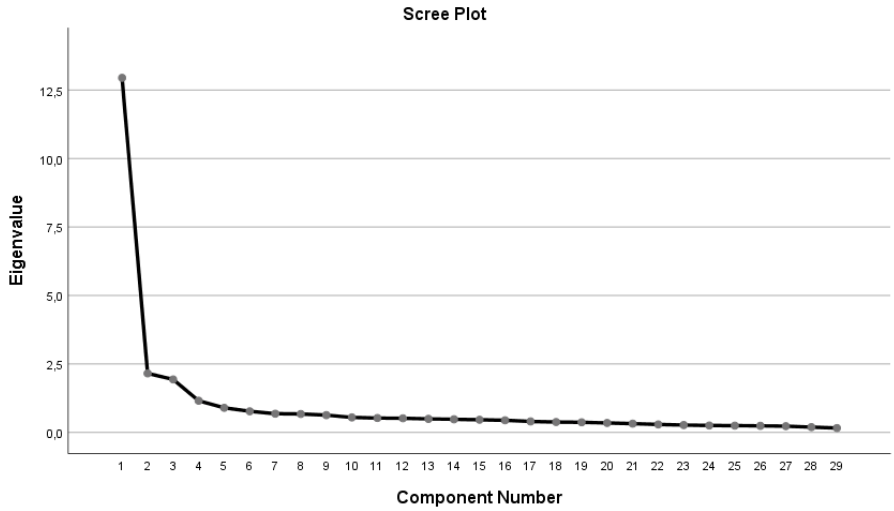
According to Table 8, four factors with eigenvalues above 1 accounted for 62.7% of the total variance. According to Özdamar (2016), it is considered sufficient for the total explained variance in the social sciences to be above 40%. The first factor had a higher eigenvalue and percentage of variance than the other factors. The first, second, third, and fourth factors accounted for 44.637%, 7.430 %, 6.671 %, and 3.984% of the total variance, respectively.

**Table 8.** *Eigenvalues.*

| Component | Initial Eigenvalues | | |
|---|---|---|---|
| | Total | % of Variance | Cumulative % |
| 1 | 12.945 | 44.637 | 44.637 |
| 2 | 2.155 | 7.430 | 52.067 |
| 3 | 1.935 | 6.671 | 58.739 |
| 4 | 1.155 | 3.984 | 62.722 |

To provide additional evidence for the factor structure of the scale, a scree plot was constructed as shown in Figure 1.

**Figure 1.** *Post-EFA scree plot.*




The scree plot and explained variance ratios indicated that the scale had a 4-factor structure. The 4-factor structure, resulting from the EFA, accounted for 62.72% of the total variance. After determining the scale's 4-factor structure, the items' factor loadings were examined. Table 9 displays the items' distribution across factors and their factor loadings.

As shown in Table 9, the factor loadings of the items ranged from .543 to .847. The distribution of items across factors was examined, and the factors were named. The naming of these factors is based on theoretical knowledge (Özdamar, 2016; Tezbaşaran, 2008). Accordingly, the factor "Participants' Learning and Use of New Knowledge and Skills (PLUNKS)" consists of 9 items (1, 2, 3, 4, 5, 6, 11, 18, 19), the factor "Organization Support (OS)" consists of 3 items (7, 8, 9), the factor " Student Learning Outcomes (SLO)" consists of 6 items (10, 12, 13, 14, 15, 16), and the factor " Participants' Reactions (PaR)" consists of 11 items (17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29). Confirmatory Factor Analysis (CFA) is conducted to verify the accuracy of the structure identified by Exploratory Factor Analysis (EFA) (Byrne, 2012). CFA was performed to confirm the 4-factor structure resulting from the EFA, and the findings of CFA are presented in Figure 2, Table 10 and 11.

**Table 9.** *Rotation matrix.*

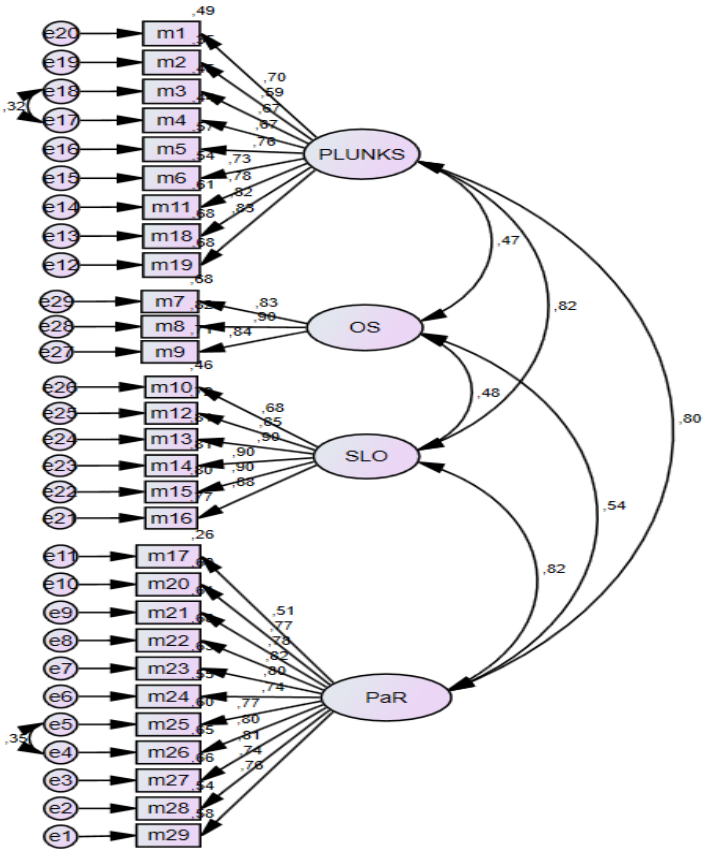
| | | Component | | | |
|---|---|---|---|---|---|
| Item No | New ranks | PaR | PLUNKS | SLO | OS |
| MG1 | M1 | | .606 | | |
| MG3 | M2 | | .696 | | |
| MG7 | M3 | | .715 | | |
| MG8 | M4 | | .679 | | |
| MG9 | M5 | | .733 | | |
| MG10 | M6 | | .724 | | |
| MG15 | M7 | | | | .814 |
| MG16 | M8 | | | | .847 |
| MG17 | M9 | | | | .830 |
| MG19 | M10 | | | .586 | |
| MG20 | M11 | | .655 | | |
| MG22 | M12 | | | .620 | |
| MG23 | M13 | | | .680 | |
| MG24 | M14 | | | .732 | |
| MG25 | M15 | | | .722 | |
| MG26 | M16 | | | .716 | |
| MG30 | M17 | .543 | | | |
| MG31 | M18 | | .625 | | |
| MG32 | M19 | | .562 | | |
| MG33 | M20 | .592 | | | |
| MG34 | M21 | .662 | | | |
| MG35 | M22 | .676 | | | |
| MG36 | M23 | .556 | | | |
| MG37 | M24 | .572 | | | |
| MG38 | M25 | .659 | | | |
| MG39 | M26 | .730 | | | |
| MG40 | M27 | .705 | | | |
| MG41 | M28 | .679 | | | |
| MG42 | M29 | .700 | | | |
| Eigen value | | 12.945 | 2.155 | 1.935 | 1.155 |
| Explained Variance | | 44.637 | 7.430 | 6.671 | 3.984 |
| Total variance | | | 62.722 | | |

In Figure 2, the interrelationships among the factors in the confirmatory factor analysis (CFA) of the professional development assessment scale are visually depicted along with the corresponding factor loadings of the individual items. The factor loadings, which represent the strength and direction of the relationships between the latent factors and observed variables, are presented in a standardized form and are shown in Figure 2. This graphical representation provides insights into the underlying structure of the scale and the extent to which each item contributes to the measurement of its respective factors.

**Table 10.** *Standardized factor loadings.*

| Item No | Std. Factor Loadings | Item No | Std. Factor Loadings | Item No | Std. Factor Loadings |
|---|---|---|---|---|---|
| M1 | 0.697 | M11 | 0.781 | M21 | 0.782 |
| M2 | 0.591 | M12 | 0.851 | M22 | 0.824 |
| M3 | 0.671 | M13 | 0.902 | M23 | 0.796 |
| M4 | 0.667 | M14 | 0.901 | M24 | 0.744 |
| M5 | 0.756 | M15 | 0.895 | M25 | 0.773 |
| M6 | 0.733 | M16 | 0.877 | M26 | 0.803 |
| M7 | 0.827 | M17 | 0.514 | M27 | 0.813 |
| M8 | 0.905 | M18 | 0.822 | M28 | 0.737 |
| M9 | 0.840 | M19 | 0.827 | M29 | 0.765 |
| M10 | 0.681 | M20 | 0.772 | | |

**Figure 2.** *The standardized factor loadings of the items in the CFA.*




The standardized factor loadings of the items included in the respective factor and the error variances of the items are shown in Table 10. Upon examination of the path diagram, it was found that the standardized factor loadings of the items under the factors were greater than 1.96 and statistically significant ($p$<.05). The standardized factor loadings of all items ranged from 0.514 to 0.905, and the error variance was found to be less than 0.630. The factor loadings of the 29 items on the scale were high, and the error variances were low; no items were removed from the scale. Goodness-of-fit indices were examined to evaluate the model as a whole, and the recommended cut-off values for goodness-of-fit indices and the fit values of the model are presented in Table 11.

**Table 11.** *Recommended criterion values for fit indices and fit values obtained from CFA.*

| Index | Excellent | Acceptable | Scale Indexes | Evaluation |
|---|---|---|---|---|
| $\chi^2/df$ | $0 \leq \chi^2/df$ <2-3 | 3<$\chi^2/df \leq 5$ | 1068.58/367=2.912 | Excellent |
| GFI | .95≤GFI≤ 1.0 | .90≤ GFI <95 | .861 | Acceptable |
| NFI | .95≤ NFI ≤1.0 | .90 ≤NFI<.95 | .905 | Acceptable |
| IFI | .95≤ TLI ≤1.0 | .90 ≤TLI<.95 | .935 | Acceptable |
| CFI | .95≤ CFI ≤1.0 | .90 ≤CFI<.95 | .935 | Acceptable |
| RMSEA | RMSEA ≤.05 | .05<RMSEA≤.08 | .063 | Acceptable |
| RMR | RMR≤.05 | .05<RMR ≤.08 | .032 | Excellent |
| SRMR | SRMR≤.05 | .05<SRMR ≤.08 | .044 | Excellent |

In the CFA, multiple fit indices were used to evaluate the model. From Table 11, it can be observed that the $\chi^2$ value divided by the degrees of freedom ($\chi^2/df$) is 2.912. When considering the other fit indices of the scale, NFI, IFI, and CFI values greater than 0.90 indicate an acceptable fit. RMSEA, RMR, and SRMR values also indicated a good fit. Overall, when considering the obtained fit values in CFA, it can be concluded that the scale, consisting of 29 items and four factors, demonstrates a good fit to the data (Bentler & Bonett, 1980; Hu &

Bentler, 1999; Jöreskog, 2004; Kline, 2016; MacCallum, Browne, & Sugawara, 1996; Schermelleh-Engel, Moosbrugger, & Müller, 2003; Schumacker & Lomax, 2010), confirming the scale structure obtained from EFA.

To provide evidence of criterion validity, the correlation between the ProDES and the Attitude Scale for Professional Development (ASPD) was calculated and presented in Table 12.

**Table 12.** *Criterion validity correlation values.*

| | | PLUNKS | OS | SLO | PaR |
|---|---|---|---|---|---|
| ASPD | *r* | .595** | .232** | .493** | .469** |
| | *p* | .000 | .002 | .000 | .000 |
| | N | 170 | 170 | 170 | 170 |

** *p*<.01

According to Table 12, the correlation values between the ASPD and the ProDES range from 0.232 to 0.595. Criterion validity refers to comparing a newly developed measurement tool with a previously validated and reliable instrument that measures the same or similar attributes (Seçer, 2015). Therefore, it can be stated that the Professional Development Evaluation Scale measures the evaluations of administrators and teachers regarding their professional development activities. Multi-group CFA analysis was also conducted to determine measurement invariance. The results are given in Table 13.

**Table 13.** *Measurement invariance results.*

| | $\chi^2$(*p*<.05) | *df* | CFI | RMSEA | TLI | SRMR | ΔCFI | ΔRMSEA |
|---|---|---|---|---|---|---|---|---|
| Configural Invariance | 1479.473 | 742 | 0.928 | 0.05 (0.046-0.054) | 0.921 | 0.049 | - | - |
| Metric Invariance | 1557.567 | 767 | 0.923 | 0.051 (0.047-0.054) | 0.918 | 0.06 | 0.005 | -0.001 |
| Scalar Invariance | 1969.936 | 792 | 0.885 | 0.061 (0.058-0.064) | 0.882 | 0.07 | 0.038 | -0.01 |

As shown in Table 13, the model-data fit in configural invariance is acceptable or excellent [$\chi^2$(742)=1479.473 (*p*<.05); CFI=0.928; RMSEA=0.05 (0.046-0.054); TLI=0.921; SRMR=0.049] was determined (Çokluk et al., 2021; Hu & Bentler, 1999; Şen, 2020) and the metric invariance stage was started. The differences between the CFI and RMSEA fit indices obtained at the configural invariance and metric invariance stages are within the criteria of ΔCFI≤.01, ΔRMSEA≤.015. After metric invariance was achieved, the scalar invariance stage was started. As a result of the analysis, the model-data fit in scalar invariance [$\chi^2$(792) = 1969.936; (*p*<.05); CFI=0.885; RMSEA=0.061 (0.058-0.064); TLI=0.882; SRMR=0.079] reached the metric invariance stage. It was observed that the difference values were worsened according to the ΔCFI≤.01, ΔRMSEA≤.015 criteria. Therefore, since the scalar invariance stage could not be achieved, strict invariance analysis was not performed.

### 3.4. Reliability Findings

One of the critical points to consider in scale development research is reducing the error rate within the total variance of the developed scale and increasing the proportion of true variance (Cohen & Swerdlik, 2015). As the error rate decreases, the reliability of the test increases (Seçer, 2015; Sönmez & Alacapınar, 2016). To achieve this, reliability evidence of the scale is reported through item-total and item-remainder correlation analyses, *t*-tests for the lower and upper group 27% groups, Cronbach's alpha, McDonald's Omega, split-half reliability, equivalent form's reliability, Guttman and Sperman-Brown coefficients, and test-retest analyses. According to these coefficients, two kinds of reliability evidence are obtained. With these coefficients, proof of reliability was obtained in terms of the internal consistency and stability of the scale.

Item-total and item-remainder correlation analyses were conducted to determine the necessity of the items in the scale and their contributions to the total score. To determine the necessity of the items in the scale and their contribution to the total score, correlation analyses between item-total, item-remainder (Table 14), and factors (Table 15) were performed.

**Table 14.** *The results of the item-total and item-remainder correlation analyses.*

| PLUNKS | | | OS | | | SLO | | | PaR | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item No | Item-total | Item-residual | Item No | Item - total | Item - residual | Item No | Item - total | Item - residual | Item No | Item - total | Item residual |
| 1 | .679** | .539** | 7 | .894** | .515** | 10 | .686** | .543** | 17 | .664** | .592** |
| 2 | .723** | .537** | 8 | .911** | .526** | 12 | .835** | .759** | 20 | .753** | .691** |
| 3 | .743** | .549** | 9 | .886** | .516** | 13 | .865** | .754** | 21 | .771** | .687** |
| 4 | .732** | .567** | | | | 14 | .850** | .694** | 22 | .793** | .695** |
| 5 | .798** | .639** | | | | 15 | .855** | .709** | 23 | .703** | .656** |
| 6 | .774** | .589** | | | | 16 | .825** | .680** | 24 | .706** | .650** |
| 11 | .785** | .662** | | | | | | | 25 | .774** | .663** |
| 18 | .759** | 663** | | | | | | | 26 | .792** | .648** |
| 19 | .765** | .725** | | | | | | | 27 | .792** | .691** |
| | | | | | | | | | 28 | .699** | .605** |
| | | | | | | | | | 29 | .736** | .623** |

** $p<.01$

**Table 15.** *Correlation analysis results between factors.*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1-PLUNKS | - | | | |
| 2-OS | .440** | - | | |
| 3-SLO | .683** | .419** | | |
| 4- PaR | .643** | .495** | .755** | - |

Correlation is significant at the 0.01 level (2-tailed), N=586

As shown in Table 14, the item-total test correlation values ranged from 0.664 to 0.911, and the item-remainder correlation values ranged from 0.515 to 0.759. Additionally, in Table 15, the interfactor correlation values ranged from 0.419 to 0.755. Correlation indicates the level and degree of relationship between items (Baykul, 2015) and/or the relationship within the dataset (Best & Kahn, 2017). When evaluating correlation coefficients, they are interpreted as follows: 0-0.29, weak or low, 0.30-0.64 moderate, 0.65-0.85 strong/high, and 0.85-1.00 very strong/very high (Ural & Kılıç, 2013). The item-total and item-remainder correlation coefficients suggest that the items in the scale are internally consistent and necessary (Cohen & Swerdlik, 2015; Özgüven, 2015). The item-total correlation values determine whether each item can be included in the total score. When an item has a low correlation coefficient, indicating a low impact on the total score, it is considered to be removed from the scale. In the item-remainder correlation, the effect of removing an item on the total score is examined, and if removing an item does not result in a significant change in the score, that item is removed (Özdamar, 2016). Based on the item-total and item-remainder correlation coefficients, it can be concluded that all items and factors in the scale demonstrate "moderate" and "high" levels of significance, indicating their relevance and importance for the scale. In other words, item-total and item-remainder correlations of 0.40 and above suggest that the items adequately measure the intended structure and effectively discriminate the intended attribute. Independent group *t*-tests should be conducted to assess the discriminant validity of scale items, distinguish between lower and upper group, or compare groups (Altunışık et al., 2004; Baker, 2016). To determine whether

the items and factors were discriminant, a 27% lower-upper group *t*-test was performed of which the findings are presented in Table 16.

**Table 16.** *27% lower and upper t-test.*

| | Lower-uppergroups | N | Mean | Sd | *t* | *df* | *p* |
|---|---|---|---|---|---|---|---|
| 1 | Lower groups | 158 | 2.32 | .824 | -14.936 | 314 | .000 |
| | Upper groups | 158 | 3.55 | .624 | | | |
| 2 | Lower groups | 158 | 2.23 | .866 | -15.229 | 314 | .000 |
| | Upper groups | 158 | 3.58 | .698 | | | |
| 3 | Lower groups | 158 | 2.47 | .720 | -15.673 | 314 | .000 |
| | Upper groups | 158 | 3.68 | .649 | | | |
| 4 | Lower groups | 158 | 2.49 | .812 | -15.935 | 314 | .000 |
| | Upper groups | 158 | 3.73 | .546 | | | |
| 5 | Lower groups | 158 | 2.46 | .803 | -18.790 | 314 | .000 |
| | Upper groups | 158 | 3.80 | .402 | | | |
| 6 | Lower groups | 158 | 2.55 | .794 | -16.806 | 314 | .000 |
| | Upper groups | 158 | 3.78 | .470 | | | |
| 7 | Lower groups | 158 | 1.90 | 1.004 | -14.760 | 314 | .000 |
| | Upper groups | 158 | 3.43 | .832 | | | |
| 8 | Lower groups | 158 | 2.06 | .942 | -15.056 | 314 | .000 |
| | Upper groups | 158 | 3.50 | .740 | | | |
| 9 | Lower groups | 158 | 2.28 | 1.040 | -14.019 | 314 | .000 |
| | Upper groups | 158 | 3.65 | .649 | | | |
| 10 | Lower groups | 158 | 2.03 | .938 | -13.932 | 314 | .000 |
| | Upper groups | 158 | 3.36 | .751 | | | |
| 11 | Lower groups | 158 | 2.31 | .756 | -20.745 | 314 | .000 |
| | Upper groups | 158 | 3.77 | .454 | | | |
| 12 | Lower groups | 158 | 2.09 | .825 | -22.727 | 314 | .000 |
| | Upper groups | 158 | 3.77 | .425 | | | |
| 13 | Lower groups | 158 | 2.01 | .740 | -24.267 | 314 | .000 |
| | Upper groups | 158 | 3.70 | .461 | | | |
| 14 | Lower groups | 158 | 1.91 | .908 | -21.934 | 314 | .000 |
| | Upper groups | 158 | 3.70 | .486 | | | |
| 15 | Lower groups | 158 | 2.06 | .879 | -19.650 | 314 | .000 |
| | Upper groups | 158 | 3.66 | .516 | | | |
| 16 | Lower groups | 158 | 2.04 | .809 | -21.465 | 314 | .000 |
| | Upper groups | 158 | 3.66 | .489 | | | |
| 17 | Lower groups | 158 | 2.01 | .971 | -17.355 | 314 | .000 |
| | Upper groups | 158 | 3.60 | .618 | | | |
| 18 | Lower groups | 158 | 2.42 | .832 | -18.796 | 314 | .000 |
| | Upper groups | 158 | 3.81 | .409 | | | |
| 19 | Lower groups | 158 | 2.30 | .720 | -23.187 | 314 | .000 |
| | Upper groups | 158 | 3.82 | .399 | | | |
| 20 | Lower groups | 158 | 1.66 | .914 | -20.662 | 314 | .000 |
| | Upper groups | 158 | 3.53 | .674 | | | |
| 21 | Lower groups | 158 | 1.83 | .925 | -19.500 | 314 | .000 |
| | Upper groups | 158 | 3.54 | .596 | | | |
| 22 | Lower groups | 158 | 1.71 | .876 | -19.268 | 314 | .000 |
| | Upper groups | 158 | 3.44 | .709 | | | |
| 23 | Lower groups | 158 | 2.26 | .783 | -19.596 | 314 | .000 |
| | Upper groups | 158 | 3.69 | .478 | | | |
| 24 | Lower groups | 158 | 2.20 | .892 | -17.956 | 314 | .000 |
| | Upper groups | 158 | 3.66 | .502 | | | |
| 25 | Lower groups | 158 | 1.70 | .907 | -18.481 | 314 | .000 |
| | Upper groups | 158 | 3.44 | .761 | | | |
| 26 | Lower groups | 158 | 1.71 | .919 | -18.343 | 314 | .000 |
| | Upper groups | 158 | 3.38 | .683 | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 27 | Lower groups | 158 | 1.92 | .896 | -19.826 | 314 | .000 |
| | Upper groups | 158 | 3.59 | .566 | | | |
| 28 | Lower groups | 158 | 2.22 | .886 | -16.943 | 314 | .000 |
| | Upper groups | 158 | 3.64 | .567 | | | |
| 29 | Lower groups | 158 | 1.92 | .944 | -16.701 | 314 | .000 |
| | Upper groups | 158 | 3.44 | .653 | | | |
| PLUNKS | Lower groups | 158 | 2.395 | .463 | -30.216 | 314 | .000 |
| | Upper groups | 158 | 3.725 | .301 | | | |
| OS | Lower groups | 158 | 2.080 | .857 | -17.207 | 314 | .000 |
| | Upper groups | 158 | 3.524 | .614 | | | |
| SLO | Lower groups | 158 | 2.023 | .532 | -31.342 | 314 | .000 |
| | Upper groups | 158 | 3.640 | .369 | | | |
| PaR | Lower groups | 158 | 1.921 | .43 | -36.132 | 314 | .000 |
| | Upper groups | 158 | 3.540 | .357 | | | |

When examining the differences in item mean scores between the lower and upper groups it can be observed that the differences in item mean scores and factors between the lower and upper groups were statistically significant at the $p$=0.001 level for all items and factors. Therefore, it can be concluded that all items and factors in the scale were discriminant. Reliability refers to the consistency of obtaining similar or identical results from the measurement tool in repeated administration. In other words, it provides an indication of the consistency of scores obtained from the measurement tool (Thorndike & Thorndike-Christ, 2017). The results of the reliability analyses conducted for this purpose are listed in Table 17.

**Table 17.** *ProDES reliability coefficients.*

| ProDES | Cronbach | McDonald's | First-Second Half | Spearman-Brown | Guttman | Split Half | Total Items |
|---|---|---|---|---|---|---|---|
| 1-PLUNKS | 0.902 | 0.904 | .829-.849 | .872 | .862 | 0.870 | 9 |
| 2-OS | 0.878 | 0.880 | .838-999 | .871 | .756 | 0.824 | 3 |
| 3-SLO | 0.900 | 0.905 | .780-.860 | .894 | .893 | 0.877 | 6 |
| 4- PaR | 0.919 | 0.920 | .859-.864 | .890 | .886 | 0.914 | 11 |

To assess the reliability of the scale, Cronbach's alpha, McDonald's Omega, split-half, equivalent forms, Guttman, and Spearman-Brown coefficients were calculated. Cronbach's alpha coefficients for the 29-item, 4-factor scale ranged from .878 to .919, McDonald's omega coefficients ranged from .880 to .920, Cronbach's alpha coefficients for the first and second halves ranged from .780 to .999, Spearman-Brown coefficients ranged from .871 to .894, and Guttman coefficients ranged from .756 to .893. Additionally, the correlation coefficients for the equivalent forms ranged from .824 to .914. On a Likert-type scale, reliability coefficients should be as close to 1 as possible (Baykul, 2015; Tezbaşaran, 2008). Reliability coefficients above $\alpha$>0.75 indicate a "high degree" of reliability (Kalaycı, 2010; Özdamar, 2016). These findings provide evidence that the scale has high overall reliability. To determine the stability and consistency reliability of the scale, it was administered twice to 170 administrators and teachers at three-week intervals. The correlation coefficients obtained from the test-retest application are presented in Table 18 and Table 19.

In Table 18, the inter-item correlation values in the test-retest application ranged from $r$=.347 to .769, in Table 19 while the inter-factor correlation values ranged from $r$=.567 to $r$=.769. The correlation values obtained from the test-retest application helped us assess the consistency of the scale over time (Kline, 2016). The stronger the correlation, the higher is the reliability (DeVellis, 2014). In this regard, the emerged correlation values indicate that the scale items and subdimensions demonstrate consistency.

**Table 18.** *Test-retest correlation values.*

| Items | PLUNKS | Items | OS | Items | SLO | Items | PaR |
|---|---|---|---|---|---|---|---|
| 1 | .514 | 7 | .578 | 10 | .418 | 17 | .446 |
| 2 | .639 | 8 | .388 | 12 | .470 | 20 | .405 |
| 3 | .565 | 9 | .516 | 13 | .484 | 21 | .657 |
| 4 | .347 | | | 14 | .355 | 22 | .703 |
| 5 | .427 | | | 15 | .402 | 23 | .424 |
| 6 | .729 | | | 16 | .367 | 24 | .769 |
| 11 | .482 | | | | | 25 | .557 |
| 18 | .360 | | | | | 26 | .470 |
| 19 | .366 | | | | | 27 | .680 |
| | | | | | | 28 | .480 |
| | | | | | | 29 | .594 |

**Table 19.** *Test-retest correlation between factors.*

| | |
|---|---|
| 1-PLUNKS | .705** |
| 2-OS | .576** |
| 3-SLO | 567** |
| 4- PaR | .769** |

** $p<.01$

## 4. DISCUSSION and CONCLUSION

Considering the positive effects of professional development on the quality of education and, ultimately, on student outcomes, it is evident how important and necessary professional development is for education systems (King, 2014). However, improving and enhancing teachers' knowledge, skills, and competencies through high-quality professional development means investing in school and student outcomes both directly and indirectly (Sancho et al., 2024).

However, a review of the literature reveals that many professional development initiatives that purport to bring about some forms of positive change are reported on by the researchers or organisations who have provided the professional development with limited evidence (that quite frequently takes the form of interview data) to substantiate the findings. Furthermore, data collection tools developed or adapted for teachers' and administrators' professional development mostly focus on either a single dimension or a specific aspect of professional development. Thus, the absence of a multidimensional data collection tool for professional development is a significant gap in the evaluation of professional development. In light of this lacuna in the research, the purpose of this research was to develop a valid and reliable scale for evaluating the professional development of administrators and teachers (Appendix).

Furthermore, the majority of scales used in research on professional development are related to teachers' attitudes towards professional development (e.g., Çakır & Horzum, 2014; Eroğlu. & Özbek, 2018, 2020; Eroğlu, 2019; Gümüş et al., 2018; Koskimäki et al., 2021; Mourão et al., 2014; Saberi & Sahragard, 2019; Shabani, et al., 2018; Torff et al., 2005; Yenen & Kılınç, 2021; Zhu, 2015). These scales, referring to teachers' attitudes towards professional development, served as an important resource for the development of the scale in the present study. In particular, the scale development process that was based on Guskey's (1986, 2000, 2002b) model of the teacher change process and Guskey's (1999, 2000, 2002a) framework for evaluating professional development, which encompasses five dimensions: Participants' learning, participants' use of new knowledge and skills, organization support and change, participants' reactions, and student learning outcomes.

According to these dimensions, an item pool of 72 items was created. The items were evaluated in terms of language and expression and 8 items were removed. The 64-item draft form was sent to 15 experts. Experts' opinions were evaluated according to the CVR and CVI criteria of the Lawshe technique, and 21 items that did not meet these criteria were removed. Content validity was ensured through the evaluation of 64 items by 15 experts, resulting in the creation of a preliminary version consisting of 43 items guided by expert opinions.

Data was collected from five different study groups along with a pilot study for the validity and reliability of the scale. Normality, KMO, and Bartlett's test values for the EFA and CFA datasets were examined, and the data were found to be suitable for factor analysis. The ANOVA Tukey Test for Nonadditivity conducted on the EFA dataset showed that total scores could not be obtained from the scale, but analysis and evaluation could be conducted using scores derived from factors. Guskey (1999, 2000) evaluates professional development at 5 levels. Since each level in Guskey's professional development evaluation model evaluates different characteristics, it supports not taking a total score from the scale. In this respect, the nonadditivity feature of the scale seems to be compatible with the theoretical background. Although the scale developed in the current study was designed as 5-dimensional, a 4-dimensional structure was obtained as a result of EFA. In Guskey's model, levels 2 and 4 are combined into one dimension. Hotelling's $T^2$ Test revealed that the items were perceived differently by the heterogeneous group. The EFA conducted on the data collected from the first study group, which consisted of 586 participants, resulted in a four-factor structure with 29 items, where the eigenvalues were above 1. Based on the literature, the factors were named "Participants' Learning and Use of New Knowledge and Skills (PLUNKS), Organization Support (OS), Student Learning Outcomes (SLO), and " Participants' Reactions (PaR)." This four-factor structure explains 62.7% of the total variance. To confirm the structure, CFA was conducted on the second study group consisting of 478 participants, and the fit indices ($\chi^2$/df ratio, NFI, IFI, CFI, GFI, RMSEA, RMR, and SRMR) reached acceptable levels. Criterion validity was established by examining the correlation between the scale and the teachers' attitudes towards the Professional Development Scale, and it was found that the correlation between these two scales was significant. The positive correlation between the two scales can also be considered an indicator of concurrent validity.

A 27% lower and upper group analysis was conducted to determine the discriminant validity of scale items. The results of the lower and upper group analyses indicated that the t-value was significant, and the discriminant values were high for all items. In other words, the item discriminant values of the ProDES indicate that it can be used to assess the professional development of administrators and teachers, as all items yielded significant differences between the lower and upper groups. The item-total and item-remainder test correlation values suggest that the scale items are important and necessary. To determine the reliability of the scale, reliability coefficients were calculated using Cronbach's alpha, McDonald's Omega, Split-half, Equivalent halves, Guttman, and Sperman-Brown methods, and it was concluded that the scale has high reliability, allowing administrators and teachers to evaluate their professional development activities/programs reliably. The final version of the scale is presented in the Appendix. In conclusion, a scale with high validity and reliability for evaluating the professional development of administrators and teachers was provided in the literature. The validated and reliable ProDES can be used by practitioners and researchers in various applications and studies involving different variables. For the scale to be applicable Türkiye and internationally, future studies should test its validity through confirmatory factor analysis and calculate reliability coefficients as evidence of measurement consistency. Educational administrators, policymakers, and researchers can use this scale to evaluate professional development activities or programs in which administrators and teachers participate.

As a result, the scale consists of 29 items and four subscales, measured on a 5-point Likert scale. The scale is evaluated as "*Strongly Disagree*=0, *Disagree*=1, *Agree*=2, *Mostly Agree*=3, and

*Strongly Agree*=4". The subscales included 9 items (1, 2, 3, 4, 5, 6, 11, 18, 19) in the " Participants' Learning and Use of New Knowledge and Skills (PLUNKS)" subscale, 3 items (7, 8, 9) in the " Organization Support (OS)" subscale, 6 items (10, 12, 13, 14, 15, 16) in the " Student Learning Outcomes (SLO)" subscale, and 11 items (17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29) in the " Participants' Reactions (PaR)" subscale. A total score could not be obtained from the scale, and there were no reverse-scored items. When comparing the subscales, their arithmetic mean was used for the evaluation. The sum of scores obtained by teachers and administrators from the subscales represents the evaluation of the quality quantity, value importance, administrative support, and contribution to students in the professional development in which they participate (degree of possessing the desired characteristic).

In general, a low score obtained by administrators and teachers from the ProDES indicates a lower level of possessing the desired characteristic, whereas a high score indicates a higher level of possessing the desired characteristic. The 30th item in the scale measures the general evaluation of teachers and administrators' professional development activities. Therefore, the 30th item was evaluated separately.

In conclusion, the scale's high internal consistency and test-retest reliability ensures that it can be used to make evidence-informed decisions that can foster more effective and supportive professional development activities. Furthermore, by identifying which professional development initiatives lead to improvements, those associated with professional development can use resources more efficiently, leading to enhanced school and system-wide improvements. Finally, the use of ProDES can also help schools and education systems to track progress over time, making ProDES an invaluable tool for continuous improvement and strategic planning across various levels of education systems.

## Acknowledgments

The author(s) would like to acknowledge the support of the Scientific and Technological Research Council of Turkey (TÜBİTAK) for funding this research through the 2219-International Postdoctoral Research Fellowship Program (Project No: 1059B192000757).

## 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**: İstanbul Sabahattin Zaim Üniversitesi, 30.05.2023-E.54478.

## Contribution of Authors

**Mustafa Özgenel:** Investigation, conception, methodology, analysis, and writing the original draft. **Martin Brown:** Investigation, conception, supervision, critical review, validation. and writing the original draft. **Joe O'Hara:** Supervision and proofreading. **Metin Özkan:** Software, data analysis, and visualization.

## Orcid

Mustafa Özgenel https://orcid.org/0000-0002-7276-4865
Martin Brown https://orcid.org/0000-0002-5436-354X
Joe O'Hara https://orcid.org/0000-0003-1956-7640
Metin Özkan https://orcid.org/0000-0002-4891-9409

## REFERENCES

Adey, P. (2004). *The professional development of teachers: Practice and theory*. Springer Science & Business Media.

Altunışık, R., Coşkun, R., Bayraktaroğlu, S., & Yıldırım, E. (2004). *Sosyal bilimlerde araştırma yöntemleri* [*Research methods in social sciences*]. Sakarya Kitabevi.

Assor, A., & Oplatka, I. (2003). Towards a comprehensive conceptual framework for understanding principals' personal-professional growth. *Journal of Educational Administration, 41*(5), 471-497.

Baker, F.B. (2017). *Madde tepki kuramının temelleri* [*Fundamentals of item response theory*] (N. Güler, Çev., Ed.). Pegem.

Barth, R.S. (1986). Principal centered professional development. *Theory Into Practice, 25*(3), 156-160.

Baykul, Y. (2015). *Eğitim ve psikolojide ölçme: Klasik test teori ve uygulaması* [*Measurement in education and psychology: Classical test theory and practice*]. ÖSYM.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Best, J.W. & Kahn, J.V. (2017). *Eğitimde araştırma yöntemleri* [*Research methods in education*] (O. Köksal. Çev. Ed.). Eğitimevi.

Blandford, S. (2004). *Professional development manual: A practical guide to planning and evaluating successful staff development*. Pearson Education.

Blandford, S. (2012). *Managing professional development in schools*. Routledge.

Bredeson, P.V. (2000) The school principal's role in teacher Professional development. *Journal of In-Service Education, 26*(2), 385-401.

Brown, T.A. & Moore, M.T. (2013). Confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361-379). Guilford Press.

Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı* [Handbook of data analysis for the social sciences]. Pegem.

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı [Factor analysis: Basic concepts and use in scale development]. *Kuram ve Uygulamada Eğitim Yönetimi, 32*, 470-483.

Byrne, B.M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming.* Routledge Taylor & Francis Group.

Campbell, A., McNamara, O., & Gilroy, P. (2004). *Practitioner research and professional development in education*. Sage.

Cohen, R.J. & Swerdlik, M.E. (2015). *Psikolojik test ve değerlendirme* [*Psychological testing and assessment*] (E. Tavşancıl, Çev. Ed.). Nobel.

Cohen, S. (2004). *Teachers' professional development and the elementary mathematics classroom: Bringing understandings to light*. Routledge.

Cole, P. (2008). *Leadership and professional learning: Forty actions leaders can take to ımprove Professional learning*. IARTV.

Çakır, Ö., & Horzum, M.B. (2014). Adaptation motivation toward web-based professional development scale and examining pre-service teachers' motivation toward web-based professional development perception in terms of different variables. *Procedia-Social and Behavioral Sciences, 131*, 144-148.

Çokluk, Ö., Şekerçioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve Lisrel uygulamaları* [*Multivariate statistics for social sciences: SPSS and Lisrel applications*]. Pegem.

Day, C. (2002). *Developing teachers: The challenges of lifelong learning*. Routledge.

Desimone, L. (2011). A Primer on professional development. *Phi Delta Kappan, 92*(6), 68-71.

DeVellis, R.F. (2014). *Ölçek geliştirme kuram ve uygulamalar* [*Scale development: theory and applications*] (T. Totan, Çev. Ed.). Nobel

Diaz-Maggioli, G. (2004). *Teacher-centered professional development*. ASCD.

Dijkstra, E.M. (2009). *Hoe professioneel is de hedendaagse onderwijsprofessional?* (What is the professionality of the contemporary educational professional?) [Unpublished master thesis]. Rijksuniversiteit Groningen.

Easton, L.B. (2008). From professional development to professional learning. *Phi Delta Kappan, 89*(10), 755-761.

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme I: Temel kavramlar ve işlemler* [*Measurement and scale development in psychology I: Basic concepts and procedures*]. Pegem.

Eroğlu, M., & Özbek, R. (2018). Development of professional development activities scale for teachers. *Journal of Current Researches on Social Sciences*, *8*(3), 185-208.

Eroğlu, M. (2019). *Öğretmenlerin mesleki gelişime katılımlarıyla, mesleki gelişime yönelik tutumları, kendi kendine öğrenmeye hazır bulunuşlukları ve destekleyici okul özellikleri arasındaki ilişkinin incelenme* [*Investigatıon of the relationship between teachers 'participation in professional development and the attitudes toward professıonal development, readıness for selfdirected learning and supportıve school characteristics*] [Unpublished doctoral dissertation]. İnönü Üniversitesi, Malatya.

Eroğlu, M., & Özbek, R. (2020). Mesleki gelişim etkinlikleri ölçeğinin uyarlanması: Geçerlik ve güvenirlik çalışması [Adaptation of professional development activities scale: Validity and reliability study]. *Turkish Studies*, *15*(4), 2611-2628.

Field, A. (2009). *Discovering statistics using SPSS*. Sage.

Fullan, M. (1994). *Change forces: Probing the depths of educational reform*. Palmer Press.

Garet, M.S., Porter, A.C., Desimone, L., Birman, B.F., & Yoon, K.S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*(4), 915-945.

George, D., & Mallery, P. (2016). *IBMM SPSS statistics step by step.* Routledge.

Guskey, T.R. (1986). Staff development and the process of teacher change. *Educational Researcher*, *15*(5), 5-12.

Guskey, T.R. (1999). *New perspectives on evaluating professional development.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Guskey, T.R. (2000). *Evaluating professional development*. Corwin press.

Guskey, T.R. (2002a). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, *8*(3/4), 381-391.

Guskey, T.R. (2002b). Does it make a difference? Evaluating professional development. *Educational Leadership*, *59*(6), 45-51.

Guskey, T.R. (2003a). *The characteristics of effective professional development: A synthesis of lists.* Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL, April 21-25, 2003).

Guskey, T.R. (2003b). Analyzing lists of the characteristics of effective professional development to promote visionary leadership. *NASSP Bulletin*, *87*(637), 4-20.

Guskey, T.R., & Yoon, K.S. (2009). What works in professional development? *Phi Delta Kappan*, *90*(7), 495-500.

Gümüş, S., Apaydın, Ç., & Bellibaş, M.Ş. (2018). Öğretmen mesleki öğrenme ölçeğinin Türkçeye uyarlanması: Geçerlik ve güvenirlik çalışması [Adaptation of teacher professional learning scale to Turkish: The validity and reliability study]. *Eğitim ve İnsani Bilimler Dergisi: Teori ve Uygulama*, *9*(17), 107-124.

Hair, J.F., Jr., Anderson, R.E., Tatham, R.L., & Black, W.C. (1998). *Multivariate data analysis* (5th ed.). Pearson Education Inc.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational Leadership. *Cambridge Journal of Education*, *33*(3), 329-352.

Helm, J.L., Castro-Schilo, L., & Oravecz, Z. (2017). Bayesian versus maximum likelihood estimation of multitrait–multimethod confirmatory factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 17-30.

Hinkin (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 2* (1), 104-121.

Hinkin, T.R., Tracey, J.B., & Enz, C.A. (1997). Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, *21*(1), 100-120.

Ho, R. (1998). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. Chapman & Hall/CRC, Boca Raton.

Hof, M.W. (2012). *Questionnaire evaluation with factor analysis and Cronbach' s Alpha: An example*. Retrieved from http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/student-papers/MHof-QuestionnaireEvaluation-2012-Cronbach-FactAnalysis.pdf

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.

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)*, 1-55.

Jöreskog, K.G. (2004). *On chi-squares for the ındependence model and fit measures in Lisrel*. http://www.ssicentral.com/lisrel/techdocs/ftb.pdf

Kalaycı, Ş. (2010). *SPSS uygulamalı çok değişkenli istatistik teknikleri* [*Multivariate statistical techniques with SPSS applications*]. Asil.

Kao, C.P., Wu, Y.T., & Tsai, C.C. (2011). Elementary school teachers' motivation toward web-based professional development, and the relationship with Internet self-efficacy and belief about web-based learning. *Teaching and Teacher Education*, *27*(2), 406-415.

Kılıç, S.F. (2022). Ölçek geliştirme sürecinde açımlayıcı faktör analizi [Exploratory factor analysis in the scale development process]. In M.Acar Güvendir & Y. Özer Özkan (Ed.), *Tüm yönleriyle ölçek geliştirme süreci* [*Scale development process in all its aspects*] (s.69-126). Pegem.

King, F. (2014). Evaluating the impact of teacher professional development: An evidence-based framework. *Professional Development in Education*, *40*(1), 89-111.

Kirkpatrick, D., & Kirkpatrick, J. (2006). *Evaluating training programs: The four levels*. Berrett-Koehler Publishers.

Kline, R.B. (2016). *Yapısal eşitlik modellemesi ve uygulaması* [Structural equation modeling and its application] (S. Şen, Çev. Ed.). Nobel.

Koskimäki, M., Mikkonen, K., Kääriäinen, M., Lähteenmäki, M.L., Kaunonen, M., Salminen, L., & Koivula, M. (2021). Development and testing of the Educators' Professional Development scale (EduProDe) for the assessment of social and health care educators' continuing professional development. *Nurse Education Today*, *98*, 104657.

Kwakman, K. (1999). *Leren van docenten tijdens de beroepsloopbaan* [*Teacher learning throughout the career*] [Unpublished doctoral dissertation]. University of Nijmegen, the Netherlands.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*(4), 563-575.

Little, J.W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis, 15*(2), 129-151.

Liu, S., Hallinger, P., & Feng, D. (2016). Supporting the professional learning of teachers in China: Does principal leadership make a difference? *Teaching and Teacher Education*, *59*, 79-91.

MacCallum, R.C., Browne, M.W., & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130.

Mourão, L., Porto, J. B., & Puente-Palacios, K. (2014). Evidence of validity of the perception of professional development scale. *Psico-USF*, *19*, 73-85.

Otrar, M., & Argın, F.S. (2015). A scale development study to determine the attitude of students towards social media. *Journal of Research in Education and teaching*, *4*(1), 391-403.

Özdamar, K. (2016). *Eğitim, sağlık ve davranış bilimlerinde ölçek ve test geliştirme yapısal eşitlik modellemesi* [*Structural equation modeling for scale and test development in education, health and behavioral sciences*]. Nisan.

Özer, N., & Beycioglu, K. (2010). The relationship between teacher professional development and burnout. *Procedia-Social and Behavioral Sciences*, *2*(2), 4928-4932.

Özgüven, İ.E. (2015). *Psikolojik testler* [Psychological tests]. Nobel.

Robson. C. (2017). *Bilimsel araştırma yöntemleri: Gerçek dünya araştırması* [*Real world research*] (Ş. Çınkır & N. Demirkasımoğlu. Çev. Ed.). Anı Publishing.

Saberi, L. & Sahragard, R. (2019). Designing and validating teachers' professional development scale: Iranian EFL contexts in focus. *International Journal of Instruction*, *12*(1), 1609-1626.

Sancho, L., Brown, M., Gardezi, S., O'Hara, J., & Rodríguez-Conde, M.J. (2024) Developing culturally responsive school leaders in Ireland and Spain. The evolving role of professional development. *Irish Educational Studies*, 1-22. https://doi.org/10.1080/03323315.2024.2334710

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.

Schumacker, R.E., & Lomax, R.G. (2010). *A beginner's guide to structural equation modeling*. Routledge Taylor ve Francis Group.

Seçer, İ. (2017). *SPSS ve LISREL ile pratik veri analizi: Analiz ve raporlaştırma* [*Practical data analysis with SPSS and LISREL: Analysis and reporting*]. Anı Publishing.

Shabani, M.B., Alibakhshi, G., Bahremand, A., & Karimi, A.R. (2018). In-service professional development scale for EFL teachers: A validation study. *The International Journal of Humanities, 25*(3), 63-78.

Soine, K.M., & Lumpe, A. (2014). Measuring characteristics of teacher professional development. *Teacher Development*, *18*(3), 303-333.

Sönmez, V., & Alacapınar, F.G. (2016). *Sosyal bilimlerde ölçme aracı geliştirme* [*Developing measurement tools in social sciences*]. Pegem.

Spillane, J.P., Healey, K., & Mesler Parise, L. (2009). School leaders' opportunities to learn: A descriptive analysis from a distributed perspective. *Educational Review, 61*(4), 407-432.

Şen, S. (2020). *Mplus ile yapısal eşitlik modellemesi uygulamaları* [Structural equation modeling applications with Mplus]. Nobel.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics (5th ed.)*. Allyn and Bacon.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics*. Allyn and Bacon.

Tatlıdil, H. (2002). *Uygulamalı çok değişkenli istatistiksel analiz* [*Applied multivariate statistical analysis*]. Akademi Matbaası.

Tavsancil, E. (2002). *Tutumlarin olculmesi ve SPSS ile veri analizi* [Measuring attitudes and data analysis with SPSS]. Nobel.

Tezbaşaran, A. (2008). *Likert tipi ölçek hazırlama kılavuzu* [*Likert type scale preparation guide*]. Üçüncü Sürüm e-Kitap. Retrieved from https://www.academia.edu/1288035/Likert_Tipi_Ölçek_Hazırlama_Kılavuzu

Thorndike, R.M. & Thorndike-Christ, T. (2017). *Psikolojide ve eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in psychology and education*] (M. Otrar, Çev. Ed.). Nobel.

Torff, B., Sessions, D., & Byrnes, K. (2005). Assessment of teachers' attitudes about professional development. *Educational and Psychological Measurement*, *65*(5), 820-830.

Ural, A., & Kılıç, İ. (2013). *Bilimsel araştırma süreci ve SPSS ile veri analizi* [*Scientific research process and data analysis with SPSS*]. Detay.

Wayne, A.J., Yoon, K.S., Zhu, P., Cronen, S., & Garet, M.S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher*, *37*(8), 469-479.

Yenen, E.T., & Kılınç, H.H. (2021). Öğretmenlerin mesleki gelişim öz yeterlikleri ölçeği geçerlik ve güvenirlik çalışması [Validity and reliability study of teachers' professional development self-efficacy scale]. *Turkish Journal of Social Research/Turkiye Sosyal Arastirmalar Dergisi*, *25*(2).

Zepeda, S.J. (2011). *Professional development: What works*. Eye on Education.

Zhu, H. (2015, November). A study on professional development scale for master of full-time education (PDSM-FE). In *2015 International Conference on Social Science, Education Management and Sports Education* (pp. 168-173). Atlantis Press.

## APPENDIX-A

| | Mesleki Gelişim Değerlendirme Ölçeği (MGDÖ) Turkish Version | Hiç katılmıyorum | Katılmıyorum | Biraz Katılıyorum | Çoğunla katılıyorum | Tamamen katılıyorum |
|---|---|---|---|---|---|---|
| 1 | Katıldığım mesleki gelişim etkinliklerinden öğrendiklerimin mesleki uygulamalarımda bir fark yarattığını düşünürüm. | 0 | 1 | 2 | 3 | 4 |
| 2 | Mesleki gelişim etkinliklerine katılmak için zaman ayırırım. | 0 | 1 | 2 | 3 | 4 |
| 3 | Öğrencilerime daha faydalı olmak için mesleki gelişim etkinliklerine katılırım. | 0 | 1 | 2 | 3 | 4 |
| 4 | Katıldığım mesleki gelişim etkinlikleri esnasında meslektaşlarımla iş birliğinde bulunarak kendimi geliştirmeye çalışırım. | 0 | 1 | 2 | 3 | 4 |
| 5 | Katıldığım mesleki gelişim etkinliklerinde edindiğim kazanımların okulda/sınıfta başarılı bir şekilde uyguladığımda mesleki gelişime yönelik tutumum olumlu yönde gelişir. | 0 | 1 | 2 | 3 | 4 |
| 6 | Katıldığım mesleki gelişim etkinliklerinin sonunda kendimi iyi hissederim. | 0 | 1 | 2 | 3 | 4 |
| 7 | Görev yaptığım okuldaki yöneticiler, okul temelli mesleki gelişim faaliyetleri düzenler. | 0 | 1 | 2 | 3 | 4 |
| 8 | Görev yaptığım okuldaki yöneticiler, eğitimcilerin mesleki gelişimini takip eder. | 0 | 1 | 2 | 3 | 4 |
| 9 | Görev yaptığım okuldaki yöneticiler, eğitimcilerin mesleki gelişim etkinliklerine katılmasını teşvik eder. | 0 | 1 | 2 | 3 | 4 |
| 10 | Katıldığım mesleki gelişim etkinlikleri öğrenci temelli istenmeyen davranışların (okulu bırakma ve disiplin vb.) azalmasını sağlar. | 0 | 1 | 2 | 3 | 4 |
| 11 | Mesleki gelişim etkinliklerinden edindiğim deneyimleri başarılı/etkili bir şekilde sınıfta/okulda uyguladığımda kendimi geliştirmeye yönelik çok daha fazla istek duyarım. | 0 | 1 | 2 | 3 | 4 |
| 12 | Katıldığım mesleki gelişim etkinlikleri öğrencileri (derse katılım, sınıf içi davranışlar ve öğrenme motivasyonları) olumlu etkiler. | 0 | 1 | 2 | 3 | 4 |
| 13 | Katıldığım mesleki gelişim etkinlikleri, öğrencilerin eğitim-öğretime yönelik tutumlarını olumlu etkiler. | 0 | 1 | 2 | 3 | 4 |
| 14 | Katıldığım mesleki gelişim etkinlikleri öğrencilerin performansını olumlu etkiler. | 0 | 1 | 2 | 3 | 4 |
| 15 | Katıldığım mesleki gelişim etkinlikleri öğrencilerin duyuşsal gelişimini destekler. | 0 | 1 | 2 | 3 | 4 |
| 16 | Katıldığım mesleki gelişim etkinlikleri öğrencilerin fiziksel/psiko-motor gelişimine katkı sağlar. | 0 | 1 | 2 | 3 | 4 |
| 17 | Katıldığım mesleki gelişim programlarının sonunda değerlendirme yapılır. | 0 | 1 | 2 | 3 | 4 |
| 18 | Mesleki gelişim etkinliklerinden edindiğim deneyimleri okulda/sınıfta başarıyla uyguladığım zaman bu tür etkinliklere katılma konusunda isteğim artar. | 0 | 1 | 2 | 3 | 4 |
| 19 | Mesleki gelişim etkinlikleri, öğretmenlerin/yöneticilerin değişim ve gelişmelere uyum sağlamasını destekler. | 0 | 1 | 2 | 3 | 4 |
| 20 | Mevcut mesleki gelişim etkinlikleri güncel mesleki ihtiyaçlarımı karşılar. | 0 | 1 | 2 | 3 | 4 |
| 21 | Mesleki gelişim etkinliklerinde amaca uygun materyaller ve araç-gereçler kullanılır. | 0 | 1 | 2 | 3 | 4 |
| 22 | Mesleki gelişim etkinlikleri eğlencelidir. | 0 | 1 | 2 | 3 | 4 |
| 23 | Mesleki gelişim etkinlerini anlamlı bulurum. | 0 | 1 | 2 | 3 | 4 |
| 24 | Mesleki gelişim etkinliklerinde hedeflenen bilgi ve becerileri kazandığımı düşünüyorum. | 0 | 1 | 2 | 3 | 4 |
| 25 | Mesleki gelişim planlayıcıları, öğretmenlerin bireysel öğrenme özelliklerini dikkate alır. | 0 | 1 | 2 | 3 | 4 |
| 26 | Mesleki gelişim planlayıcıları, öğretmenlerin mesleki gelişimle ilgili yaşadıkları problemlere göre düzenleme yaparlar. | 0 | 1 | 2 | 3 | 4 |
| 27 | Mesleki gelişim eğitimcileri, yeni bilgi ve becerileri sınıfa/okula nasıl aktaracağım konusunda fikirler sunar. | 0 | 1 | 2 | 3 | 4 |
| 28 | Katıldığım mesleki gelişim etkinlerinde görev alan eğitimciler alanlarında yetkin kişilerdir. | 0 | 1 | 2 | 3 | 4 |
| 29 | Katıldığım mesleki gelişim etkinliklerinde ortaya çıkan sorunlar hızlı bir şekilde çözülür. | 0 | 1 | 2 | 3 | 4 |
| 30 | Lütfen şu ana kadar katıldığınız mesleki gelişim etkinliklerini genel anlamda değerlendirerek 0-100 arasında bir puan vererek değerlendiriniz: | ..... | | | | |

## APPENDIX-B

| | Professional Development Evaluation Scale (ProDES)<br>English Version | Strongly disagree | Disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| 1 | I think what I learned from the professional development activities I've attended made a difference in my professional practice. | 0 | 1 | 2 | 3 | 4 |
| 2 | I spare time to attend professional development activities. | 0 | 1 | 2 | 3 | 4 |
| 3 | I participate in professional development activities to be more beneficial to my students. | 0 | 1 | 2 | 3 | 4 |
| 4 | I try to improve myself through cooperation with my colleagues in the professional development activities I attend. | 0 | 1 | 2 | 3 | 4 |
| 5 | My attitude towards professional development develops in a positive way when I successfully apply the gains I've achieved in the professional development activities at school/class. | 0 | 1 | 2 | 3 | 4 |
| 6 | I feel good at the end of the professional development activities I attend. | 0 | 1 | 2 | 3 | 4 |
| 7 | The administrators at my school organize school-based professional development activities. | 0 | 1 | 2 | 3 | 4 |
| 8 | The administrators at my school follow the professional development of teachers. | 0 | 1 | 2 | 3 | 4 |
| 9 | The administrators at my school encourage educators to participate in professional development activities. | 0 | 1 | 2 | 3 | 4 |
| 10 | The professional development activities I attend contribute in reducing undesirable student behaviors (dropping out of school and discipline, etc.) in my school. | 0 | 1 | 2 | 3 | 4 |
| 11 | When I successfully/effectively apply the experiences I gained from professional development activities in the classroom/school, I feel much more willing to improve myself. | 0 | 1 | 2 | 3 | 4 |
| 12 | The professional development activities I attend positively affect my students (increased class participation, desirable classroom behaviors and learning motivations). | 0 | 1 | 2 | 3 | 4 |
| 13 | The professional development activities I attend positively affect students' attitudes towards teaching and learning. | 0 | 1 | 2 | 3 | 4 |
| 14 | The professional development activities I attend positively impact students' performance. | 0 | 1 | 2 | 3 | 4 |
| 15 | The professional development activities I attend support students' emotional development. | 0 | 1 | 2 | 3 | 4 |
| 16 | The professional development activities I attend contribute to students' physical/psycho-motor development. | 0 | 1 | 2 | 3 | 4 |
| 17 | Evaluation is made at the end of the professional development programs I attended. | 0 | 1 | 2 | 3 | 4 |
| 18 | My desire to participate in such activities increases when I successfully apply the experiences I gained from professional development activities at school/classroom. | 0 | 1 | 2 | 3 | 4 |
| 19 | Professional development activities support teachers/administrators to adapt to changes and developments. | 0 | 1 | 2 | 3 | 4 |
| 20 | Present professional development activities meet my current professional needs. | 0 | 1 | 2 | 3 | 4 |
| 21 | Appropriate materials and tools are used in professional development activities. | 0 | 1 | 2 | 3 | 4 |
| 22 | Professional development activities are fun. | 0 | 1 | 2 | 3 | 4 |
| 23 | I find professional development activities meaningful. | 0 | 1 | 2 | 3 | 4 |
| 24 | I think I've acquired the knowledge and skills targeted in professional development activities. | 0 | 1 | 2 | 3 | 4 |
| 25 | Professional development planners take into account the individual learning characteristics of teachers. | 0 | 1 | 2 | 3 | 4 |
| 26 | Professional development planners make adjustments according to the problems teachers/administrators experience with professional development. | 0 | 1 | 2 | 3 | 4 |
| 27 | Professional development trainers provide insights into the ways through which the transfer of novel knowledge and skills to the classroom/school take place. | 0 | 1 | 2 | 3 | 4 |
| 28 | The trainers involved in the professional development activities I attend are qualified individuals in their fields. | 0 | 1 | 2 | 3 | 4 |
| 29 | Problems that arise in the professional development activities I participate in are resolved quickly. | 0 | 1 | 2 | 3 | 4 |
| 30 | Please evaluate the professional development activities that you have participated in so far and give a score between 0-100, considering them in a general sense: | ..... | | | | |

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 758–773
https://doi.org/10.21449/ijate.1495346
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Over-education rates and predictors of entry-level jobs in Türkiye

**Sevgi Ernas**[1*]

[1]Ankara University, Faculty of Education Science, Department of Elementary Education, Ankara, Türkiye

**ARTICLE HISTORY**

*Received: June 03, 2024*

*Accepted: Aug. 26, 2024*

**Keywords:**

Education mismatch,
Overeducation,
Overskilling,
Education-employment relationship,
Theories of education economics.

**Abstract:** This research aims to determine the proportion of overeducated individuals with higher education levels compared to their colleagues who are graduates of associate, undergraduate, and postgraduate education but work at the same status in entry-level jobs. Overeducation rates in entry-level jobs in Türkiye were determined using the Turkish Statistical Institute (TUIK) Household Labor Force Surveys (2014-2019) microdata set. The job analyst measure was used to determine the rate of overeducation. Logistic regression data analysis was conducted to classify the variables that predict the state of being overeducated with the TUIK 2019 Household Labor Force Survey. According to the findings, overeducation rates increased gradually over the years by 8.02% in 2014, 8.98% in 2015, 9.78% in 2016, 10.43% in 2017, 11.00% in 2018, and 12.5% in 2019. For the state of being overeducated, various demographic variables were analyzed and predicted, such as income, age, region, gender, ISCED, marital status, firm size, place of work, additional job searches, ISCO 08 classification, and employment status.

## 1. INTRODUCTION

After World War II, although the expansion of the education sector and professions significantly slowed down after 1970, the increase in the educated workforce accelerated. This situation indicates that from the 1970s to the early 1980s, professions were educationally rising (Clogg & Shockey, 1984), and an increase in the duration and level of schooling among workers in the United States was observed (Halaby, 1994). In the 1970s, an increase in the number of graduates in the United States and the rising demand for graduates in the workforce led to the emergence of the phenomenon of "overeducation" (Berg, 1970; Freeman, 1976). This phenomenon still holds true (International Labor Organization [ILO], 2019; Kurnaz, 2015). The phenomenon of "overeducation" occurs when the number of educated individuals increases and the educational level on the supply side of the labor market exceeds the level demanded for employment. When the labor market cannot absorb the increasing supply of educated labor, i.e., when there is an imbalance between supply and demand, educated individuals are forced to accept jobs that do not match their education qualifications, thus falling into an "overeducated" situation (Büchel, 2001).

*CONTACT: Sevgi ERNAS ernasevgi@gmail.com Ankara University, Faculty of Education Science, Department of Elementary Education, Ankara, Türkiye



**e-ISSN: 2148-7456**

### 1.1. Overeducation

"Overeducation" refers to the mismatch between the educational attainment of the workforce and the level of education required for jobs (Rumberger, 1981). Often, there is a discrepancy between the qualifications offered by the education system and those demanded by the labor market. Although the term "qualification" is used to denote the attainment of or exceeding defined or definable minimum criteria, the criteria for required qualifications for a job are debated (Ünal, 1996). The presence of a mismatch in the labor market is commonly addressed as horizontal and vertical according to labor market theories (ILO, 2019; Kurnaz, 2015; Quintini, 2011a). Horizontal mismatch refers to the situation where knowledge and skills acquired through education are not utilized, whereas vertical mismatch refers to individuals working in jobs below their qualifications (ILO, 2019; Quintini, 2011a). The discrepancy between the education levels of individuals in the labor market and the jobs they perform is termed a qualification mismatch or an educational mismatch. Mismatch occurs when individuals have higher or lower educational qualifications than those required by their jobs, resulting in "overeducation" or "undereducation" (ILO, 2019; Kurnaz, 2015; Quintini, 2011a).

Experimental studies on "overeducation" are categorized into three main categories. First, there are skill and education requirements for each job, accepted by job analysts and countries such as the United States, the Netherlands, and Portugal (Chevalier, 2003). Second, self-assessment of educational requirements by employees is defined (Green, Mcintosh & Vignoles, 1999). Third, education distribution is calculated for each occupation, with deviations from the mean (Verdugo & Verdugo, 1988) or mode (Mendes de Oliviera, Santos & Kiker, 2000) and some specific values (usually a standard deviation) (Chevalier & Walker, 2001). Research analyzing the relationship between education and income has shown that individuals who are overeducated for their jobs face significant wage penalties compared to those with similar educational backgrounds working in jobs that match their qualifications (Chevalier & Walker, 2001). In international studies on overeducation, the impact of overeducation on earnings has been associated with issues such as job satisfaction and job mobility (Delaney et al., 2020; McGuinness, 2006; McGuinness et al., 2018; Pouliakas, 2012; Quintini, 2011b). Experimental studies have been conducted on how earnings are shaped when there is a mismatch between the educational level of the employed person and the educational level required by the job. These studies show income losses for individuals who are overeducated for their jobs. Conversely, the incomes of individuals who are undereducated for their jobs tend to be higher than those of individuals with the same level of education (Sicherman, 1991). Mendes et al. (2000) found that while overeducated workers should earn more than their equally educated but not overeducated colleagues, they earn less than their adequately educated colleagues.

### 1.2. Overeducation in the Context of Educational Economic Theories

The fundamental principle of Human Capital Theory is that the skills acquired through education represent human capital, which employers value and leads to increased productivity. This productivity is also rewarded with higher wages (Becker, 1975). The theory also demonstrates that education and training are investments. The basic approach of the theory is that short-term expenses can provide “cash flow” in the long term. As with other investment plans, cost-benefit analyses, such as using the internal rate of return, can be performed (Psacharopoulos, 1987). Human Capital Theory primarily explains the supply side of the labor market and does not address job requirements on the demand side (Hartog & Oosterbeek, 1988). Jobs and job requirements are considered consistent elements (homogeneous factors), and these variables are not included in the factors of earnings and matching. Human Capital Theory does not accept mismatched matches and asserts that individuals will reach the most suitable position in the labor market. Any mismatch situation existing in the labor market is also considered temporary within the context of Human Capital Theory (Desjardins & Rubenson, 2011).

It is stated that the low wages of overeducated individuals are due to variables not considered in the measurements (Kucel, 2011). However, it is accepted that the fundamental argument of Human Capital Theory—that earnings increase as the level of education increases—is inconsistent due to the phenomenon of overeducation (Dolton & Vignoles, 2000). According to Human Capital Theory, individuals with lower levels of education are more likely to be unemployed than those with higher education levels. According to the theory, the failure of the education system to respond at the same pace to changes in the labor market and the lack of new graduates who can adapt to new jobs emerging as a result of technological developments are among the causes of unemployment (Kurnaz, 2015). Consequently, wages are always aligned with the marginal product of an individual worker, which is determined by the level of human capital accumulated through formal education or on-the-job training (Quintini, 2011b). In this context, as firms adjust their production processes to fully utilize individuals' human capital or as this situation persists, educational mismatches can be eliminated in the short term.

According to the Screening Hypothesis, the formal recognition of an individual's qualifications through diplomas and certificates offered by the education system during job placement can lead to the phenomenon of overeducation due to qualification inflation and the exclusion effect (Desjardins & Rubenson, 2011). The increase in the number of highly educated individuals is among the reasons for qualification inflation, as it reduces the importance, distinctiveness, and prestige of having high educational qualifications and thus the selection feature (Kurnaz, 2015). Qualification inflation also indicates that as the number of highly educated individuals increases, the level of qualification decreases. Employers will not be able to fully utilize the qualifications and skills of the workforce unless they adapt their production technologies to the workforce, leading to a loss of earnings for individuals as labor productivity does not increase. Ultimately, situations of "over-education" and "over-skilling" will emerge, where the qualifications and skills possessed by employees in the labor market are not utilized, resulting in a potential loss of value in investments made through education (Desjardins & Rubenson, 2011).

According to the Queue Hypothesis, qualification mismatch is considered a permanent phenomenon in the labor market. Additionally, according to the Queue Hypothesis, there is no wage return for overeducation, i.e., having an education above the job requirements. According to the hypothesis, wages are determined entirely based on the educational qualifications required for the job (Quintini, 2011a). The Queue Hypothesis characterizes a market where individuals compete for job opportunities based on their relative education costs rather than competition based on wages determined by their human capital (McGuinness, 2006). According to the Job Competition Model on which the Queue Hypothesis is based, individuals with inadequate education and skills can succeed in competition for qualified jobs and earn higher incomes (Desjardins & Rubenson, 2011; McGuinness et al., 2018).

Employers raising the qualifications at the hiring stage direct individuals forming the supply to receive more education than the job requires to obtain the desired job or to advance their job positions. Entry-level jobs, as the most visible part of labor markets, are viewed by employers as a tool for temporarily selecting candidates for highly qualified positions (Aksoy, 1998). "Over-educated" individuals, having received education above the level required for the job they perform, will accept lower jobs to find employment, thus forming the subject of this research problem in entry-level jobs. This study aims to determine the ratios of "over-educated" individuals who graduated from associate, bachelor's, and postgraduate education levels, who work in entry-level jobs, having higher education levels compared to their colleagues in the same status. It addresses this issue in the context of the education-employment relationship. To achieve this aim, the following questions were asked:

1. What are the levels of "over-education" in entry-level jobs in Türkiye, and do they change over the years?

2. What are the predictors of being over-educated in entry-level jobs in Türkiye?

## 2. METHOD

The investigation was based on a detailed analysis of overeducation rates and predictors of entry-level jobs in Türkiye. The context in which the study took place is described in the research model/design, sampling, data collection tool, and data analysis.

### 2.1. Research Model

Quantitative methodology was employed in this study. The proportions of highly educated individuals, including those with associate, bachelor, and postgraduate degrees, employed at entry-level positions in Türkiye were determined. Additionally, variables predicting highly educated individuals employed at entry-level positions in Türkiye were identified. The research adopted a survey and a correlational research design. Survey research designs involve researchers collecting information from a sample group selected from a population or the entire population to explain the attitudes, opinions, behaviors, and characteristics of individuals in that population (Creswell, 2017). Using data from the Turkish Statistical Institute's Household Labor Force Survey between 2014 and 2019, the proportions of overeducated individuals were determined over the years through a longitudinal survey design called "panel studies." Panel studies are longitudinal survey designs that examine the same group of people over a specified period (Creswell, 2017).

To address the second aim of the study, variables predicting overeducation were identified using data from the 2019 Turkish Statistical Institute's Household Labor Force Survey. This section of the study employed a predictive design based on correlational research. The goal of predictive research design is to identify variables that forecast specific outcomes and criteria. (Creswell, 2017).

### 2.2. Sampling and Data Collecting Tool

In this study, the entire sample from the Turkish Statistical Institute's Household Labor Force Survey between 2014 and 2019 was used to determine the number and proportion of overeducated individuals employed in entry-level jobs. The data for this research were obtained from the "Micro Data Set of Household Labor Force Survey" conducted by the Turkish Statistical Institute. The Household Labor Force Survey covers the years 2014-2019. Access to these data was obtained electronically through official correspondence between Ankara University's Institute of Educational Sciences and the Turkish Statistical Institute.

### 2.3. Data Analysis

In this section, the data analysis methods used in the research process are presented according to the sequence of research questions. For the first research question, data from the Turkish Statistical Institute's Household Labor Force Survey were analyzed to examine the overeducation status of individuals employed in entry-level jobs. A matching matrix was applied to determine the overeducation rate. The matching matrix was constructed in four stages. In the first stage, only the employed individuals from the panel dataset were considered; in the second stage, those employed in entry-level jobs were identified; in the third stage, graduates with bachelor's and postgraduate degrees were identified; and in the fourth stage, those employed in entry-level jobs with bachelor's and postgraduate degrees, i.e., overeducated individuals employed in entry-level jobs, were determined. ISCO 08 codes were utilized to define entry-level jobs and identify graduates in these jobs. The ISCO 08 occupational classifications published in 2012 were used for occupational classifications in the dataset. The proportions of overeducated individuals were determined by years through percentage and frequency analyses and are presented in tables.

For the second research question, multiple logistic regression analysis was conducted to determine the variables that predicted whether employees are overeducated or not. Logistic

regression analysis is a statistical technique used when the dependent variable is binary or multinomial. This analysis predicts the probability of belonging to a certain class of dependent variables and models the relationship between independent variables and dependent variables. The logistic regression model transforms probabilities using a function called the logit function, and predictions are made based on this logit transformation (Menard, 2002). As the independent variable considered in this research is the overeducation status of employees, binary logistic regression analyses were performed. The predictors included employee gender, place of residence, age, nationality, type of employment, working hours, and income. While continuous variables were directly included in the analysis, categorical variables with more than two subgroups were coded as dummy variables and included in the analysis. The standard (enter) method was used because all variables were included simultaneously in the analysis (Field, 2018).

Before analysis, the assumptions of logistic regression were tested. An effort was made to achieve a participant size of 10-15 times the number of variables to ensure the adequacy of the sample size. Because 366.556 participants were reached in this research, this assumption was met. Variance inflation factor (VIF) values were examined to determine whether multicollinearity existed among the predictor variables. The VIF values for all variables were found to be less than 10, indicating no multicollinearity issues. Standard residuals were examined to identify univariate outliers, with variables outside the range of 3 to +3 considered outliers. Cook's distance and Mahalanobis distance coefficients were calculated to identify multivariate outliers. In this context, observations with Cook's distance greater than 1 and Mahalanobis distance coefficients statistically significant ($p$ <0.05) were excluded from the analysis (161 observations). Model data fit was tested using the Hosmer-Lemeshow test before the main analysis, and it was decided that the model fit was adequate ($\chi^2$ = 56.893, $p$> .05). The findings are presented in tables.

## 3. RESULTS

This section presents the overeducation rates derived from the analysis of data from the Turkish Statistical Institute's (TUIK) Household Labor Force Survey between 2014 and 2019, as well as the variables predicting overeducation from the analysis of the 2019 Household Labor Force Survey data.

### 3.1. Overeducation Rates Among Entry-Level Workers

To determine the overeducation rates among entry-level workers, the education levels, employment statuses, and occupations according to the International Standard Classification of Occupations (ISCO) of participants in the TUI Household Labour Force Survey were examined. By analyzing the dataset from 2014 to 2019, the overeducation rate among entry-level workers in Türkiye was determined. Before determining the overeducation rates, the distribution of variables that determine overeducation across years is shown. Table 1 provides the distribution of demographic information regarding the education and employment status of the workforce according to the TUIK 2014, 2015, 2016, 2017, 2018, and 2019 Household Labour Force Survey. These demographic details are the variables used to determine the overeducation rate.

To determine the rate of overeducation, the job analyst method was employed. This method, which is used to create occupational dictionaries, relies on evaluations by professional job analysts tasked with measuring educational requirements by occupation. The job analysis method has been used by Thurow and Lucas (1972), Hartog and Oosterbeek (1988), Kiker and Santos (1991) in Portugal, and Hartog (2000) in the Netherlands. Rumberger (1987) analyzed the relationship between educational mismatch and earnings using this classification. It is also possible to define over- and under-education using the International Standard Classification of Education (ISCED) for large occupational groups and the International Standard Classification of Occupations (ISCO) for classifying by education level. For instance, ISCO classifies top

executives and managers as having a higher education level (ISCED 5-6) (McGuinness et al., 2018).

To determine the overeducation rates, a four-stage matrix process was conducted using the job analyst method. This model is based on systematic evaluation by job analysts of the necessary education level and type for occupations classified by education level. The job analysis method relies on evaluations by professional job analysts tasked with measuring educational requirements by occupation. Table 1 presents the distribution of overeducated entry-level workers over the years.

**Table 1.** *Rates of overeducation in entry-level jobs by year (2014-2019).*

| Year | Sample Size | Employed | | Entry-Level Job Worker | | Associate, Bachelor's, and Postgraduate Graduates | | Over-Educated in Entry-Level Jobs | Over-Education Rates |
|---|---|---|---|---|---|---|---|---|---|
| | *N* | *f* | % | *f* | % | *f* | % | *f* | % |
| 2014 | 393.822 | 174.287 | 44.2 | 117.797 | 67.5 | 43.660 | 11.0 | 9.459 | 8.0 |
| 2015 | 389.035 | 174.452 | 44.8 | 116.148 | 66.5 | 46.060 | 19.8 | 10.437 | 8.9 |
| 2016 | 380.709 | 171.402 | 45.0 | 112.571 | 65.6 | 48.861 | 12.8 | 11.013 | 9.7 |
| 2017 | 378.691 | 171.152 | 45.2 | 112.589 | 65.7 | 51.003 | 12.4 | 11.745 | 10.4 |
| 2018 | 374.179 | 170.240 | 45.5 | 111.352 | 65.4 | 52.905 | 14.1 | 12.249 | 11.0 |
| 2019 | 366.556 | 161.300 | 44.0 | 104.354 | 64.7 | 55.477 | 15.1 | 12.689 | 12.1 |

**Source:** Created by the author based on data from TÜİK (Turkish Statistical Institute).

Table 1, which shows the rates of overeducated entry-level workers, indicates that the employment rate was 44.2% in 2014, with some partial increases over the years, although the lowest employment rate was 44% in 2019. The rate of entry-level workers decreased gradually from 67.5% in 2014 to 64.7% in 2019. However, the percentage of associate, undergraduate, and postgraduate graduates increased from 11.0% in 2014 to 15.1% in 2019. The increase in the educational levels of individuals on the supply side also affects the educational levels of employed persons on the demand side.

### 3.2. Variables Predicting Overeducation Among Entry-Level Workers

This section identifies the variables that predict the overeducation status of entry-level workers based on the TUIK 2019 Household Force Surveys. Descriptive analysis and logistic regression results of variables predicting overeducation, supported by the literature, are presented here.

#### 3.2.1. *Descriptive statistics of variables predicting overeducation*

The descriptive statistics of variables predicting overeducation include personal information, working style, earnings, statistical region classification, firm characteristics, International Standard Classification of Occupations (ISCO 08), and International Standard Classification of Education (ISCED-F). Descriptive statistics are categorized into two categories: entry-level workers and overeducated entry-level workers. Table 2 shows the distribution of personal information among entry-level workers who participated in the TUIK 2019 Household Labour Force Survey.

When examining the gender distribution of entry-level workers in Table 2, 67.1% are male and 32.9% are female, while 70.7% are male and 29.3% are female. This indicates that the proportion of male workers is higher than that of female workers among overeducated entry-level workers. Regarding the marital status of entry-level workers, the highest proportion is married individuals at 74.1%, followed by never married individuals at 21.6%, divorced individuals at 2.7%, and widowed individuals at 1.6%. Among overeducated entry-level workers, 59.1% are married, 37.8% have never married, 2.8% are divorced, and 0.3% are

widowed. This finding highlights that married individuals are the majority of overeducated entry-level workers.

**Table 2.** *Distribution of entry-level workers by personal information (2019).*

| Personal Information | | Over-Educated | | Total | |
|---|---|---|---|---|---|
| | | *f* | % | *f* | % |
| Gender | Female | 4.176 | 32.9 | 30.619 | 29.3 |
| | Male | 8.513 | 67.1 | 73.735 | 70.7 |
| | Total | 12689 | 100 | 104.354 | 100.0 |
| Marital Status | Never Married | 4.801 | 37.8 | 22.532 | 21.6 |
| | Married | 7.498 | 59.1 | 77.361 | 74.1 |
| | Divorced | 352 | 2.8 | 2.790 | 2.7 |
| | Widowed | 38 | 0.3 | 1.671 | 1.6 |
| | Total | 12.689 | 100.0 | 104.354 | 100.0 |
| Age | 15-24 Years Old | 1.716 | 13.52 | 14.278 | 13.7 |
| | 25-34 Years Old | 5.565 | 43.85 | 21.567 | 20.7 |
| | 35-44 Years Old | 3.223 | 25.39 | 26.493 | 25.4 |
| | 45-54 Years Old | 1.474 | 11.61 | 22.819 | 21.9 |
| | 55 and over | 711 | 5.60 | 19.197 | 18.4 |
| | Total | 12.689 | 100 | 10.4354 | 100.0 |
| Place of Residence | Provincial Center | 3.814 | 30.1 | 11.978 | 11.5 |
| | Distict Center | 2.889 | 22.8 | 14.006 | 13.4 |
| | Town or Village | 378 | 3.0 | 7.249 | 6.9 |
| | Total | 7.081 | 55.8 | 33.233 | 31.8 |
| | Unspecified | 5.608 | 44.2 | 71.121 | 68.2 |
| | Total | 12.689 | 100.0 | 104.354 | 100.0 |

**Source:** Created by the author based on data from TÜİK (Turkish Statistical Institute).

Examining the age distribution of entry-level workers in Table 2, the highest proportion is workers aged 35-44 at 25.4%, followed by 45-54 at 21.9%, 25-34 at 20.7%, 55 and over at 18.4%, and 15-24 at 13.7%. Among overeducated entry-level workers, 43.8% are aged 25-34, 25.3% are aged 35-44, 13.5% are aged 15-24, and 5.6% are aged 55 and over. This indicates that the highest proportion of overeducated entry-level workers is in the 25-34 age group.

The distribution of entry-level workers by place of residence in Table 2 shows that the majority live in district centers, whereas the distribution of overeducated workers by place of residence indicates that the highest proportion, 30.1%, live in provincial centers. This finding considers that overeducated entry-level workers are more likely to live in provincial centers because job opportunities are predominantly available in these areas. Table 3 presents the distribution of entry-level workers according to employment information.

Table 3 shows the distribution of entry-level workers by earnings according to the TUIK Household Labour Force Survey. According to the distribution of earnings in Table 3, the highest proportion of entry-level workers, 30.7%, earned between 0 and 2.020 TL, followed by 22.2% earning between 2.020-4.000 TL, 3.9% earning between 4.001-7.000 TL, and 0.2% earning over 7.000 TL. For over-educated workers, 37.2% earn between 2.020-4.000 TL, 25.0% earn between 0-2.020 TL, 19.9% earn between 4.001-7.000 TL, and 0.8% earn over 7.000 TL. In summary, the highest proportion of over-educated workers, 37.2%, earned between 2.020-4.000 TL.

**Table 3.** *Distribution of earnings for entry-level workers participating in the TÜIK household labor force survey (2019).*

| Rank | Income Groups | Over-Education | | Total | |
|---|---|---|---|---|---|
| | | *f* | % | *f* | % |
| 1 | 0-2.020 TL | 3.175 | 25.0 | 32.082 | 30.7 |
| 2 | 2.020- 4000 TL | 4.723 | 37.2 | 23.160 | 22.2 |
| 3 | 4.001-7.000 TL | 2.535 | 19.9 | 4.046 | 3.9 |
| 4 | 7000 TL and over | 102 | 0.8 | 169 | 0.2 |
| 5 | Total | 10.535 | 83.0 | 59.457 | 57.0 |
| 6 | Unspecified | 2.154 | 16.9 | 44.897 | 43.0 |
| | Total | 12.689 | 100 | 104.354 | 100.0 |

**Source:** Created by the author based on data from TÜİK (Turkish Statistical Institute).

### 3.2.2. *Variables predicting overeducation among entry-level workers*

Logistic regression analysis was performed to identify variables that accurately classified the overeducation status of individuals after examining the assumptions required for logistic regression analysis in the dataset used in the research. Logistic regression included personal information (gender, marital status, age, place of residence), employment information (SGK registration status, job status, number of employees in the workplace, job finding method, working style, job continuity, side job status, job search status, lifelong participation in activities), ISCO 08, income, ISCED classification, and region classification as variables to classify overeducation. Initially, the "forward LR" method was used to include variables in the analysis. Variables that did not significantly contribute to the model were excluded. According to Field (2009), if the exclusion of an independent variable results in a significant difference in model fit, the variable is retained in the model. Subsequently, the analysis was repeated using the "enter method" with the significant variables. The initial model obtained with significant variables had a 2LL value of 26.651.254, which is a likelihood value similar to the sum of squares that indicates how well the maximum likelihood estimation fits (Çokluk et al., 2010).

Regarding the initial model, the constant term, its error, the Wald statistic (154.89), the degrees of freedom (1) of the Wald statistic, the significance level ($p$=.000), and the exponential logistic regression coefficient (Exp(β)= 1.19) are given. The significant outcome of the error chi-square statistic ($\chi\beta02$ = 8029.020, $p\leq.05$) for predictor variables not included in the initial model suggests that adding these predictor variables to the model would increase its predictive power. In the initial model without independent variables, the program classified all participants as overeducated, resulting in a correct classification percentage of 54.5%. The omnibus test results for the intended model after logistic regression analysis are provided in Table 4.

**Table 4.** *Omnibus test of the model coefficients.*

| Step | | Chi-square ($\chi^2$) | *df* | *p* |
|---|---|---|---|---|
| 1 | Step | 9902.260 | 51 | .000 |
| | Blok | 9902.260 | 51 | .000 |
| | Model | 9902.260 | 51 | .000 |

Upon examining Table 4, the $p$-value for the chi-square statistic was found to be significant. This indicates the presence of a relationship between the dependent and predictor variables. The result of the Hosmer and Lemeshow test, calculated when the independent variables are included in the model, is presented in Table 5.

**Table 5.** *Hosmer–Lemeshow test*.

| Step | Chi-square ($\chi^2$) | *df* | *p* |
|---|---|---|---|
| 1 | 56.893 | 8 | .060 |

The non-significant Hosmer and Lemeshow test statistic ($\chi^2$ = 56.893, $p$>.05) in Table 5 indicates that the model-data fit is adequate and that there is a relationship between the predictor and predicted variables. This implies that the model predictions do not significantly differ from the observed cases. The final classification status of the dependent variable after logistic regression analysis is given in Table 6.

**Table 6.** Summary of the targeted model *with the predictor variables*.

| | -2LL | Cox and Snell $R^2$ | Nagelkerke $R^2$ |
|---|---|---|---|
| Step 1 | 17116662 | .389 | .5206 |

In Table 6, the 2LL value of the intended model with predictor variables is 16. The initial model's 2LL value was 26.651.254, and the decrease to 16.748.993 in the intended model signifies a significant improvement in model fit. The 2LL difference of 9.902.261 indicates improvement due to predictor variables (Çokluk, 2010). Additionally, the Cox and Snell $R^2$value shows that predictor variables explain 40.1% of the variance in overeducation status. The Nagelkerke $R^2$ value is 52%, indicating the proportion of variance explained by the logistic model, where higher values correspond to better model fit (Hair et al., 2019). Table 7 lists the predictor variables not included in the initial model.

It has been determined that the variables of gender (Wald: 24.39, $p$<0.05), age (Wald: 128.63, $p$<0.05), ISCEDDF (Wald: 2576.93, $p$<0.05), marital status (Wald: 277.29, $p$<0.05), number of employees (Wald: 22.54, $p$<0.05), ISCO08 (Wald: 817.19, $p$<0.05), additional employment status (Wald: 11.26, $p$<0.05), and income (Wald: 294.81, $p$<0.05) statistically significantly predict the likelihood of being overeducated. However, it has been found that working style (Wald: 1.205, $p$>0.05) and job continuity (Wald: 0.899, $p$>0.05) are not significant predictors of overeducation.

Considering the gender variable, men are 0.78 times less likely to be overeducated compared to women (B=-0.235, ExpB=0.790). In other words, women are 1.26 times more likely to be overeducated than men. A one-unit increase in age increases the likelihood of being overeducated by 1.01 times (B=0.013, ExpB=1.013). Individuals included in the ISCEDF_K3 field are 0.002 times less likely to be overeducated compared to others (B=-0.063, ExpB=0.002). In other words, individuals in this occupational group are 500 times more likely not to be overeducated compared to other occupational groups. Divorced individuals are 0.28 times less likely to be overeducated compared to others (married, single, widowed) (B=-1.268, ExpB=0.281). In other words, divorced individuals are 3.56 times more likely not to be overeducated compared to others.

Additionally, individuals working in workplaces with 50 or more employees are 1.24 times more likely to be overeducated compared to those with fewer employees (B=0.212, ExpB=1.236).

Individuals in the ISCO08K10 occupational group are 0.076 times less likely to be overeducated compared to individuals in other occupational groups (B=-2.573, ExpB=0.076). In other words, individuals in this occupational group are approximately 13.15 times more likely not to be overeducated compared to others. Individuals with additional employment are 1.50 times more likely to be overeducated compared to those without additional employment (B=0.402, ExpB=1.495).

Finally, individuals with an income level of 3 or higher are 3.43 times more likely to be overeducated compared to those with lower income levels (B=1.232, ExpB=3.426). The final classification status of the dependent variable after logistic regression analysis is provided in Table 7.

**Table 7.** *Final classification status of dependent variables after logistic regression analysis.*

| Observed Value | Predicted Value | | Correct Classification Percentage |
|---|---|---|---|
| | NotOver-Education | Over-education | |
| Not OverEducated | 6.810 | 1.992 | 77.7 |
| OverEducated | 2.107 | 8.428 | 80.0 |
| Total Correct Classification Percentage | | | 78.8 |

In Table 7, the logistic regression analysis shows that 78.8% of the overeducation status was accurately classified. Of the 8.802 individuals not overeducated, 6.810 were correctly classified, whereas 1.992 were incorrectly classified as overeducated. Of the 10.535 overeducated individuals, 8.428 were correctly classified as overeducated, whereas 2.107 were incorrectly classified as not overeducated, with a correct classification rate of 80%.

While the overall classification percentage in the model without the inclusion of variables (null model) was 54.5%, it increased to 78.8% in the model with the inclusion of variables. In this case, it can be stated that the variables contributed to the classification power of the model and strengthened it." In other words, these results clearly demonstrate that the model performs better and increases its classification accuracy when independent variables are included.

## 4. DISCUSSION and CONCLUSION

This section discusses the overeducation rates among entry-level workers between 2014 and 2019 and the variables predicting overeducation.

### 4.1. Discussion and Conclusion on Over-Education Rates Among Entry-Level Workers Between 2014 and 2019

The job analyst method, which involves evaluations by professional job analysts tasked with measuring educational requirements by occupation, was used to determine overeducation rates. This method has been employed by Thurow and Lucas (1972), Hartog and Oosterbeek(1988)., Kiker and Santos (1991) in Portugal, and Oosterbeek and Webbink (1996, as cited in Hartog, 2000) in the Netherlands. Rumberger (1987) analyzed the relationship between educational mismatch and earnings using this classification. The International Standard Classification of Education (ISCED) can define over- and under-education for large occupational groups, whereas the International Standard Classification of Occupations (ISCO) can be used to classify by education level. For instance, ISCO classifies top executives and managers as having a higher education level (ISCED 5-6) (McGuinness et al., 2018).

According to data from the TUIK Household Labour Force Survey, the number of employed individuals showed an increasing trend until 2018 but decreased in 2019. During the same period, the number of entry-level workers decreased, whereas the rates and numbers of overeducated associate, undergraduate, and postgraduate graduates increased. According to the TUIK Household Labour Force Survey, the employment rate slightly decreased from 44.25% in 2014 to 44% in 2019, despite some increases in certain years. Parallel to these data, the rate of entry-level workers decreased gradually from 67.0% in 2014 to 64.70% in 2019. Examining the schooling rates on the supply side, the rate of associate, undergraduate, and postgraduate graduates increased from 11.02% in 2014 to 15.13% in 2019. Accordingly, the number of higher education graduates on the supply side increased. The increase in the rates of associate, undergraduate, and postgraduate graduates has also raised the education levels of individuals eligible for employment on the demand side. The rise in education levels on the supply side,

without adequately meeting the demand for jobs requiring higher education, has led to a growth in overeducation rates. Overeducation rates among entry-level workers increased gradually from 8.02% in 2014 to 8.98% in 2015, 9.78% in 2016, 10.43% in 2017, 11.00% in 2018, and 12.5% in 2019.

According to OECD (2019) data, the education levels of the workforce have increased, leading to a higher number of highly educated workers for jobs. The overeducation rates showed a gradual increase from 8.0% in 2014, 8.9% in 2015, 9.7% in 2016, 10.4% in 2017, 11.0% in 2018, and 12.5% in 2019. The increasing overeducation rates over the years indicate a future imbalance in the labor market. Overeducation rates have increased in developed countries due to rising higher education participation rates in recent years (Delaney et al., 2020). The increase in higher education participation rates raises the growth rate of the workforce and overeducation rates, while also increasing unemployment rates, negatively impacting returns to education (Groot & Maassen van den Brink, 2000).

McGuinness et al.'s (2018) review of 98 overeducation studies based on approximately 40 high-income countries found that the overeducation rate remained around 18% in many European Union countries from 2003 to 2013, with an average overeducation rate of 24%. Compared with other countries' data, Handei et al., (2016) found that the overeducation rate in the STEP sample was 22.3% in North Macedonia and 70.1% in Vietnam, with an average rate of 36%. These rates are much higher than those in developed labor markets. According to the International Labour Organization's (ILO, 2019) School to Work Transition Survey (SWTS), the overeducation rate among young people was 16%, with an inter-country average of 47% in low- and middle-income countries. Comparing these data with the overeducation rate in Türkiye, it can be said that the overeducation rate in Türkiye is lower.

### 4.2. Discussion and Conclusion on Variables Predicting Overeducation

According to the 2019 TUIK Household Force Survey, the number of overeducated male workers in entry-level jobs is higher than that of their female counterparts. The majority of entry-level workers in the labor market are male. Among the overeducated individuals in entry-level jobs, the highest proportion are married, followed by never-married, divorced, and widowed individuals. In terms of age distribution, the over-educated entry-level workers are primarily aged 25-34, followed by those aged 35-44, 15-24, 45-54, and 55 and over. Overeducated entry-level workers predominantly reside in provincial centers, followed by districts and villages. The most overeducated individuals hold associate or undergraduate degrees, followed by postgraduate or doctoral degrees. Examining the work locations of overeducated individuals, the majority work in the private sector, followed by the public sector and other organizations (foundations, associations, cooperatives, political parties, NGOs, international organizations, and embassies). In the public sector, entry-level workers are more likely to match the required education levels for their jobs.

According to Frank (1978), married women are more likely to be overeducated because they tend to seek jobs near their spouses' workplaces. Evidence also suggests that married women are more over-educated than their spouses (McGoldrick & Robst, 1996). García-Mainar et al., (2014) attribute this to women traditionally occupying female-dominated occupations, which often require lower education and skill levels. The lower number of over-educated female workers in entry-level jobs in Türkiye differs from the literature. This can be attributed to the lower labor force participation rate of women compared with men in Türkiye, as shown in Table 3. The European Commission's (2019) Türkiye Report on Employment and Social Policy highlights that the primary source of inequality and gender discrimination is the low labor force participation rate of Turkish women. The report also indicates a significant gap (38%) between the employment, labor force participation, and unemployment rates of men and women.

In a study examining the relationship between skill mismatch, educational participation, and structural changes in employment in Sub-Saharan African countries, Sparreboom and Gomis

(2015) found that overeducation increases with age, and women are more likely to be overeducated or undereducated than men. This finding is similar to the lower overeducation rate among those aged 55 years and above.

According to the 2019 TUIK Household Labour Force Survey, most overeducated entry-level workers are registered with the Social Security Institution (SGK). Overeducated entry-level workers are predominantly paid employees, followed by employers, self-employed individuals, and unpaid family workers. The majority of overeducated workers are employed in workplaces with 50 or more employees, followed by those with 10 or fewer, 20-49, and 11-19 employees. Thus, overeducated individuals are mostly employed in large-scale workplaces.

Overeducated individuals primarily found jobs through their own efforts, relatives, friends, acquaintances, the Turkish Employment Agency, and private employment offices. Most overeducated individuals work full-time and in permanent jobs and generally do not have side jobs. Most overeducated individuals are not actively looking for a new job. A very low proportion of overeducated individuals participate in lifelong learning activities. The regional classification (IBBS) of entry-level workers' distribution shows that the regions with the highest number of overeducated workers are, in order: the Aegean Region, Western Anatolia Region, Mediterranean Region, Istanbul Region, Eastern Marmara Region, Western Black Sea Region, Western Marmara Region, Central Eastern Anatolia Region, Southeastern Anatolia Region, Central Anatolia Region, Eastern Black Sea Region, and Northeastern Anatolia Region.

Franzen (2006) found that graduates who found jobs through communication networks or direct employer communication were more likely to find jobs requiring qualifications than those who used formal job search methods. This finding contradicts the distribution of job search methods among overeducated individuals, where the largest proportion (25.1%) answered "by my own means." This discrepancy can explain the high proportion of individuals (65%) who did not respond to the relevant question. Additionally, job searching through official institutions and career offices reduces overeducation due to the information asymmetry between applicants and employers (Carroll & Tani, 2015). The low proportion of overeducated individuals who found jobs through the Turkish Employment Agency is consistent with Carroll and Tani's (2015) findings.

According to the ISCED-F classification of education and training fields, overeducated individuals are predominantly educated in business and management, engineering and engineering operations, social sciences and behavioral sciences, education, personal services, and security services. The rates of over-educated individuals in other education and training fields are as follows: information and communication technologies, agriculture, forestry, and fisheries, manufacturing and processing, humanities, architecture and construction, health, arts, physical sciences, languages, and welfare (social services), law, occupational health and transport services, journalism and information, biology and environmental science, mathematics and statistics, and veterinary medicine. According to the TUIK Household Labour Force Survey, the most common occupations of overeducated entry-level workers according to the International Standard Classification of Occupations (ISCO 08) are sales workers, followed by protection services workers, general office clerks, keyboard clerks, and numerical and material recording clerks.

Logistic regression analysis identified 11 variables predicting the overeducation status of entry-level workers: income, age, region, gender, ISCED, marital status, firm size, work location, side job search status, ISCO 08 classification, and job status. "In the initial classification of the logistic regression analysis, the baseline classification accuracy for the dependent variable, overeducation, was 54.50%, while the final classification accuracy was correctly predicted at 81.2%. Budria and Moro-Egido (2018), using data from the European Skills and Jobs Survey,

found that overeducation rates were higher among part-time workers. This finding differs from that of overeducated entry-level workers in Türkiye.

The literature indicates that overeducated individuals experience negative earnings outcomes compared with their well-matched peers (Kucel, 2011). Many studies on the impact of overeducation on income show that overeducated individuals experience earnings losses. According to the "Hunger and Poverty Threshold" survey by TURK-IS (Confederation of Turkish Trade Unions) (2019), the poverty line for a family of four was 6.733 TL. Considering that most overeducated workers earn between 2.020 and 4.000 TL, they are likely living at or below the poverty line, indicating significant earnings losses.

According to an ILO (2019) study, overeducated individuals with side jobs have lower wages, less job satisfaction, and earn more additional income than their colleagues. Individuals have managed to increase their productivity levels (Leuven & Oosterbeek, 2011; McGuinness, 2006; Quintini, 2011b). Literature on the relationship between education and income indicates that overeducated individuals experience significant earnings losses compared to individuals working in the same jobs (Delaney et al., 2020). McGuinness et al., (2018), in their study examining 98 overeducation studies based on approximately 40 high-income countries, found evidence of income losses among overeducated individuals. The inclusion of income as a predictor of overeducation is consistent with the literature. The identified variables predicting overeducation are equivalent to the findings of research in the literature.

Although the overeducation rate in Türkiye is lower than that in other countries, the increase rates over the years and the accompanying overkilling issue indicate that it will become a problem for the labor market in the future. The continued education of individuals, especially when not matched by the supply side, is one of the problematic elements of the labor market. Universities should review their programs to ensure that the skills imparted align with labor market needs. In this way, graduates will possess the necessary qualifications during the implementation phase in the job market. Additionally, longitudinal studies on overeducation rates can help take preventive measures as the rates increase. Therefore, it is essential to continue research on these topics to develop policies related to these phenomena in universities, relevant ministries, and labor market sectors. Research can also be conducted on other occupational classifications beyond entry-level jobs, which is a limitation of this study.

## Acknowledgments

The paper was derived from the PhD thesis completed at Ankara University Graduate School of Educational Sciences under the supervision of Prof. Dr. Hasan Hüseyin Aksoy.

## Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number**: Ankara University, Ethics Committee, 2021/16.

## Orcid

Sevgi Ernas https://orcid.org/0000-0003-1213-7285

## REFERENCES

Aksoy, H.H. (1998). Relationship between education and employment: how do employes use educational indicators in hiring? *Journal of Interdisciplinary Education, 3*(1). https://files.eric.ed.gov/fulltext/ED426207.pdf

Atasoy, D. (2001). *Lojistik regresyon analizinin ve bir uygulaması.* [*Examination of logistic regression analysis and an application*] [Master's dissertation, Cumhuriyet University].

Becker, G.S. (1975). *Human capital: a theoretical and emprical analysis with special reference to education*. NBER.

Berg, I. (1970). *Education and jobs: The great training robbery*. Praeger.
Budra, S., & Moro-Egido, A.I. (2018). Qualication and skill mismatches: Europe from a cross-national perspective. *Cuadernos Económicos de ICE, 95*, 151-188. https://doi.org/10.32796/cice.2018.95.6646
Büchel, F. (2001). *Overqualification: reasons, measurement issues and typological affinity to unemployment.* Office for Official Publications of the European Communities. https://www.cedefop.europa.eu/files/3008EN244Buechel.pdf
Carroll, D., & Tani, M. (2015). Job search as a determinant of graduate overeducation: Evidence from Australia. *Education Economics, 23*(5), 631–644. https://ftp.iza.org/dp7202.pdf
Chevalier, A., & Walker, I. (2001). Further results on the returns to education in the UK. (I. Walker, N. Westergard-Nielsenand & C. Harmon (Eds.), *Education and earnings in europe: a cross-country analysis of returns to education*. (pp.302-330). Edward Elgar. http://eprints.lse.ac.uk/19277/
Chevalier, A. (2003). Measuring overeducation. *Economica, 70*(279), 509-531. https://doi.org/10.1111/1468-0335.t01-1-00296
Clogg, C.C., & Shockey, J.W. (1984). Mismatch between occupation and schooling: prevalence measure, recent trends and demographic Analysis. *Demography, 21*(2), 235-257. https://doi.org/10.2307/2061042.
Cohn, E., & Ng, Y.C. (2000). Incidence and wage effects of overschooling and in Hong Kong. *Economics of Education Review, 19*, 159-168. https://www.sciencedirect.com/science/article/pii/S0272-7757(99)00006-0
Creswell, J.W. (2017). *Eğitim araştırmaları: nicel ve nitel araştırmanın planlanması yürütülmesi ve değerlendirilmesi*. [*Educational research planning, conducting and evaluating quantitative and qualitative research*] (Çev. H. Ekşi). Eğitim Danışmanlığı ve Araştırmaları Merkezi.
Çokluk, Ö. (2010). Lojistik regresyon analizi: Kavram ve uygulama, [Logistic regression analysis: Concept and application] Educational Sciences in Theory and Practice, *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1357-1407.
Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları [Multivariate statistics for social sciences: SPSS and LISREL applications]*. Pegem Akademi Yayıncılık.
Delaney, J., McGuinness, S., Pouliakas, K., & Redmond, P. (2020). Educational expansion and overeducation of young graduates: a comparative analysis of 30 European countries. *Oxford Review of Education*, *46*(1), 10-29. https://doi.org/10.1080/03054985.2019.1687433
Desjardins, R., & Rubenson, K. *(2011). An Analysis of Skill Mismatch Using Direct Measures of Skills, OECD Education Working Papers. OECD Publishing.* https://doi.org/10.1787/5kg3nh9h52g5-en
Dolton, P.J., & Vignoles, A. (2000). The incidence and effects of overeducation in the U.K. graduate labor market, *Economics of Education Review, 19*, 179–198.
Dunn, D.S. (2001). *Statistics and data analysis for the behavioral sciences.* McGraw Companies.
Field, A. (2009). *Discovering statistics using SPSS (Third Ed.)*. SAGE Publications.
Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). Sage Publications.
Frank, R.H. (1978). Why women earn less: The theory and estimation of differential over qualification, *American Economic Revie*w, *68*(3*)*, 360- 373.
Franzen, A. (2006). Social networks and labor market outcomes: The non-monetary benefits of social capital. *European Sociological Review, 22(*4), 353-368. https://doi.org/10.1093/esr/jcl001
Freeman, R.B. (1976). *The overeducated American*. Academic Press.
Freeman, R.B. (1979). Why is there a youth labor market problem? *NBER Working Paper Series, 365*. 1-32. https://doi.org/10.3386/w0365

García-Mainar, I., García-Martín, G., & Montuenga, V. (2014). Overeducation and gender occupational differences in Spain. *Social Indicators Research, 124*(3), 807–833. https://doi.org/10.1007/s11205-014-0811-7

Green, F., Mcıntosh, S., & Vignoles, A. (1999). *Overeducation and Skills: Clarifying the Concepts*. *Centre for Economic Performance Discussion Paper*, London: School of Economics and Political Science. http://cep.lse.ac.uk

Groot, W., & Maassen Van Den Brink, H. (2000). Overeducation in the labor market: A meta analysis. *Economics of Education Review, 19,* 149-58. http://www.sciencedirect.com/science/article/pii/S0272-7757(99)00057-6

Hair, F.J., Black, C.W., Babin, B.J., & Anderson, E.R. (2019). *Multivariate data analysis* (Eighth Ed.). Cengage Learning.

Halaby, C.N. (1994). Overeducation and skill mismatch. *Sociology of Education, 67(1),* 47-59. https://doi.org/10.2307/2112749

Handel., M.J., Valerio, A., & Sanchez Puerta, M.L. (2016). Accounting for mismatch in low- and middle-income countries: Measurement, magnitudes, and explanations. *Directions in Development* World Bank Group. http://documents.worldbank.org/curated/en/837391472639964572/Accounting-for-mismatch-in-low-and-middle-income-countries-measurement-magnitudes-and-explanations

Hartog, J., & Oosterbeek, H. (1988). Education, allocation and earnings in the Netherlands: 0verschooling? *Economics of Education Review, 7(2),* 185-194. https://doi.org/10.1016/0272-7757(88)90043-X

Hartog, J. (2000). Overeducation and earnings: Where are we and where should we go? *Economics of Education Review, 19*, 131–147. https://doi.org/10.1016/S0272-7757(99)00050-3

International Labour Organization. (2019). Skills and jobs mismatches in low and middle income countries. GILO.

Kalaycı, Ş. (2014). *SPSS uygulamalı çok değişkenli istatistik teknikleri*] [*Multivariate statistical techniques with SPSS*]. Asil Yayıncılık.

Kiker, B.F., & Santos, M.C. (1991). Human capital and earnings in Portugal. *Economics of Education Review, 10* (3), 187–203. https://doi.org/10.1016/0272-7757(91)90043-O

Kucel, A. (2011) Literature survey of the incidence of overeducation: a sociological approach. *Revista Espanola de Investigaciones Sociológicas, 134*, 125-142. http://www.reis.cis.es/REIS/PDF/Reis_134_061302519925436.pdf

Kurnaz, I. (2015). *İşgücü piyasalarında uyumsuz eşleşme ve aşırı eğitimlilik olgusu: Türkiye'nin ilk 500 firması, Ankara ili örneği*. [*Overeducatıon and mısmatch ın the labor market: the case of top 500 largest ındustrıal organızatıons of Turkey, Ankara provınce*] [Doctoral dissertation, Gazi University].

Leuven, E., & Oosterbeek, H. (2011). Overeducation and mismatch in the labor market, *IZA Discussion Paper Series, 5523*, 1- 53. https://docs.iza.org/dp5523.pdf

Long, J.S. (1997). *Regression models for categorical and limited dependent variables.* Sage Publications.

McGuinness, S. (2006). Overeducation in the labor market. *Journal Econonic of Surveys, 20,* 387–418. https://doi.org/10.1111/j.0950-0804.2006.00284

McGuinness, S., Pouliakas, K., & Redmond, P. (2018) Skills mismatch: concepts, measurement and policy approaches, *Journal of Economic Surveys, 32(4),* 985-1015. https://doi.org/10.1111/joes.12254

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage Publications.

Mendes De Oliviera, M., Santos, M., & Kiker, B. (2000). The role of human capital and technological change in overeducation. *Economics of Education Review, 19*, 199–206. https://doi.org/10.1016/S0272-7757(99)00020-5

Mertler, C.A., & Reinhart, R.V. (2017). *Advanced and multivariate statistical methods practical application and interpretation.* Routledge.

Nieto, S., & Ramos, R. (2017). Overeducation, skills and wage penalty: evidence for Spain using PIAAC data. *Social Indicators Research, 134*(1), 219-236. https://doi.org/10.1007/s11205-016-1423-1

Pallant, J. (2005). *SPSS survival manual: a step by step guide to data analysis using SPSS.* Allen & Unwin Publications.

Pouliakas, K. (2012). The skill mismatch challenge in Europe. In: European Commission (Eds.). *Employment and social developments in Europe* (pp. 351- 394). Publications Office,

Quintini, G. (2011a). *Over-qualified or under-Skilled: A review of existing literature*. OECD Social, Employment and Migration Working Papers. OECD Publishing. https://doi.org/10.1787/1815199X

Quintini, G. (2011b). *Right for the Job: Over-Qualifiedor Under-Skilled?* France: OECD Social, Employment and Migration Working Papers. https://doi.org/10.1787/5kg59fcz3tkd-en

Rumberger, R. (1981). *Overeducation in the U.S. labor market*. Praeger.

Rumberger, R. (1987). The impact of surplus schooling on productivity and earnings. *Journal of Human Resources, 22(1),* 24–50. https://doi.org/10.2307/145998

Sicherman, N. (1991). Overeducation in the labor market. *Journal of Labor Economics, 9*(2), *101-122* https://www.jstor.org/stable/2535236?seq=1/subjects.

Stephenson, B. (2008). *Binary response and logistic regression analysis.* http://pages.stat.wisc.edu/~mchung/teaching/MIA/reading/GLM.logistic.Rpackage.pdf

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics (Fifth Ed.).* Pearson Publications.

Thurow, J., & Lucas, R. (1972). *The American distribution of income: a structural problem*. U.S Government Printing of Fice. https://www.jec.senate.gov/reports/92nd%20Congress/The%20American%20Distribution%20of%20Income%20%20A%20Structural%20Problem%20(546).pdf

Tinsley, H.E.A., & Brown, S.D. (2000). *Handbook of applied multivariate statistics and mathematical modeling.* Elsevier Science & Technology Books.

Ünal, I.L. (1996). *Eğitim ve yetiştirme ekonomisi*. [*Economics of education and training*]. Epar Yayınları.

Verdugo, R., & Verdugo, N. (1988). The ımpact of surplus schooling on earnings: some additional findings. *The Journal of Human Resources, 24*(4), 629-643. https://doi.org/10.2307/145998

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 774–786
https://doi.org/10.21449/ijate.1499099
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Examination of differential item functioning in PISA through univariate and multivariate matching differential item functioning

**Ahmet Yıldırım**[1*], **Nizamettin Koç**[2]

[1]Ankara Hacı Bayram Veli University, Faculty of Literature, Department of Psychology, Ankara, Türkiye
[2]Ankara University, Faculty of Educational Sciences, Department of Educational Sciences, Retired, Ankara, Türkiye

**ARTICLE HISTORY**

*Received: June 10, 2024*

*Accepted: Sep. 02, 2024*

**Keywords:**

Differential item functioning,

Multivariate matching,

Purified matching variable.

**Abstract:** The present research aims to examine whether the questions in the Program for the International Student Assessment (PISA) 2009 reading literacy instrument display differential item functioning (DIF) among the Turkish, French, and American samples based on univariate and multivariate matching techniques before and after the total score, which is the matching variable, is purified of the items flagged with DIF. The study is a correlational survey model research, and the participants of the study consist of 4459 Turkish, French, and American students who took booklets 1, 3, 4, and 6 in the PISA 2009 reading literacy measure. Univariate and multivariate (bivariate, trivariate, and quadrivariate) DIF analyses were performed through logistic regression before and after purifying the matching variable off the items displaying DIF. Literature was used to detect extra matching variables, and multiple linear regression analysis was carried out. As a result of the analyses, it was discovered that using extra matching variables apart from the total score reduces type I errors. It was also concluded that the exclusion of DIF items (removal of items with DIF) while calculating the total score led to variation in the number of questions detected as DIF and DIF levels of the items, although it did not yield consistent results.

## 1. INTRODUCTION

Adapting measures developed in linguistic community for use in different communities is a practice frequently used in recent years (Allalouf, Hambleton & Sireci, 1999). The translation of the Binet-Simon Intelligence Test from the original language to the source language can be considered one of the oldest samples of this (Hambleton, 1993; Hambleton & Patsula, 1999). Cross-cultural studies require adaptation of measures and administration in various communities (Van de Vijver & Tanzer, 2004). However, ensuring that the measured structure is equivalent across all cultures is crucial for making meaningful interpretations (Braun & Harkness, 2005; Gierl, 2000).

Recently, there has been a noticeable increase in intercultural evaluation studies conducted internationally, as well as in the number of countries participating in these studies. For example, a total of 65 countries and non-members of the Organization for Economic Co-operation and

*CONTACT: Ahmet YILDIRIM ahmet-yildirim@hbv.edu.tr Ankara Hacı Bayram Veli University, Faculty of Literature, Department of Psychology, Ankara, Türkiye



e-ISSN: 2148-7456

Development (OECD) participated in PISA (Program for International Student Assessment PISA) in 2012, in which Turkey also participated. Similarly, 63 countries got involved in Trends in International Mathematics and Science Study (TIMSS) in 2011 (International Association for the Evaluation of Educational Achievement, 2012). Considering that the countries participating in these studies and the people living in these countries differ in terms of ethnicity, language, and many other variables (Sireci and Rios, 2013) the necessity of adapting the tests developed within the scope of international studies to the language and culture of the participating countries emerges.

In adaptation studies, it is an important validity issue that the instruments adapted are not comparable with the original tests (Arffman, 2010; Ercikan et al., 2004; Perrone, 2006; Sireci & Allalouf, 2003). Because when the scores obtained from the tests are not comparable, it becomes difficult to make comparable interpretations based on the scores of the individuals taking the test in the cross-cultural studies (American Educational Research Association, 2014). PISA is one of the crosscultural studies administered in many different countries. Wealthier countries tend to participate in PISA as they have an assessment culture and also would like to see the trends in their educational system based on time. However, economically disadvantaged countries also started to show interest in large-scale international research so that they can see improvement in their education system. Currently, lower-middle-income countries such as Georgia and Indonesia; and upper-middle-income countries like Bulgaria and Brazil have participated in PISA administrations. As a result, PISA has a huge coverage in terms of participation (Organisation for Economic Co-operation and Development, 2015). The aim of PISA is to determine the competencies of 15-year-old students in three main areas: (a) reading skills, (b) mathematics, and (c) science literacy. Regardless of the construct measured by the test, there are basically two factors that affect the equivalence of measurement instruments used in international studies such as PISA: (1) translation, (2) culture (Gradshtein, Mead & Gibby, 2010).

As the utilization of tests in making important education-related decisions increases and legal issues concerning the use of tests arise, differential item functioning (DIF) and item bias may become an important problem in the evaluation of test validity (Hambleton, Clauser, Mazor & Jones, 1993). Bias causes systematic errors that deform the outcomes acquired from the measures and the evaluation based on these findings (Gierl, Rogers & Klinger, 1999). As testing and testing practices have come to public attention in recent years, test publishers and experts who use tests have to provide evidence that the tests they use and publish are not biased against minorities and are invariant for all participant groups (Hambleton et al., 1991).

Recently, DIF analyses have been frequently utilised to detect items that are not comparable across different communities (Allalouf et al, 1999; Allalouf & Sireci, 1998; Gierl et al., 1999; Gierl & Khaliq, 2000). DIF analyses are used to determine whether the test items function similarly across different groups (Hambleton et al., 1993; Sireci & Swaminathan, 1996; Zumbo, 1999; Zumbo, 2007).

DIF refers to the psychometric difference in how a question functions for two different groups. In other words, DIF can be defined as the distinction in performance between the groups compared concerning the relevant item (Allalouf et al., 1999; Dorans & Holland, 1993). DIF happens when a question in a test works inequivalently for various groups (Clauser & Mazor, 1998; Furlow et al., 2009). The reasons that make it necessary to conduct DIF studies are (Zumbo, 2007): (1) ensuring equity and fairness in assessment and evaluation, (2) Eliminating possible threats for validity, (3) Examining the equivalence of translated tests.

In DIF analyses, individuals in different groups are matched based on a matching variable and contrasted with regard to their performance on items (Camilli, 1992). The determination of a valid and justifiable matching variable is important for obtaining precise results in DIF analyses (Gierl et al., 2000). In DIF analyses, the sum of the item scores (endogenous variable) is usually

employed as the matching variable (Hambleton et al., 1993; Sireci & Rios, 2013). How valid and reliable such matching will be is a question that needs to be answered. It is suggested that matching should be based on an external variable with previously established validity (Gierl, 2004). Unfortunately, such a variable may not always be available (Clauser & Mazor, 1998). The use of additional matching variables should be considered when other variables are thought to be related to the construct or affect individuals' performance on the construct being measured (Sireci & Rios, 2013).

When the secondary factors that lead to the emergence of DIF are elements of the construct assessed by the measure and are consciously measured, these factors are referred to as auxiliary factors. However, when these factors are measured even though they are not components of the construct assessed by the instrument, they are called confounding factors (Boughton et al., 2000; Camilli, 1992; Gierl & Khaliq, 2000). DIF led by auxiliary factors is called benign DIF, while DIF led by confounding factors is called malignant DIF (Boughton et al., 2000; Gierl, 2004). DIF analyses based on multivariate matching provide a better understanding of the causes of DIF and reduce the likelihood of making type I errors (Roussos & Stout, 1996). Within the framework of DIF, the type I error is the detection of an item with DIF when in reality the item does not display DIF (Jodoin, 1999). Determining a reliable and error-free matching variable is critical for obtaining accurate results in DIF studies. Whether the matching variable should be purified of the items with DIF is an important question to be answered in DIF analyses (Sireci & Rios, 2013). The involvement of DIF items in the total score while calculating the matching variable calls into question the appropriateness of the matching variable (Gierl et al., 2000). When conducting DIF analyses, the matching variable needs to be purified. In other words, items labeled as DIF should be discarded and the total score should be recomputed. This recomputed total score is employed as the matching variable for the second logistic regression analysis (Zumbo, 1999). French and Maller (2007) state that the involvement of DIF items in the total score in DIF detection may lead to errors. To control these errors, researchers (French & Maller, 2007; Gierl et al., 2000; Khalid & Glas, 2013; Zumbo, 1999) argue that the total score, which is the main matching variable, should be purified. According to Lee and Geisinger (2016), the purification of the matching variable involves the exclusion of items defined as DIF in the initial DIF analysis when calculating the total score, to put it another way, the use of only non-DIF items when calculating the matching variable (when calculating the total score). Two approaches are adopted in the purification of the matching variable. One of these is the two-stage purification approach and the other is the iterative purification approach. When a single DIF study is conducted to exclude DIF items from the calculation of the matching variable, it is referred to as the two-stage purification approach. If iterative DIF analyses are performed until no items are identified as DIF, it is known as the iterative purification approach (Lee & Geisinger, 2016).

As PISA is an intercultural evaluation study, both English and French versions of all measures used within the scope of PISA are developed, and these tools are sent to the participating countries for adaptation procedures. The two forms of the test are developed in parallel and in this way, it is planned to minimize cultural dependency. As a result of the adaptation, the various language forms of the test are considered to be the same. However, it needs to be demonstrated whether this is the case in reality. Moreover, in DIF studies conducted on items of international tests such as PISA, individuals are usually matched using a single matching variable (total scores) and analyses are conducted in this way. In addition, DIF analyses are conducted without purifying the total score which is the matching variable of the items with DIF. Considering that other variables such as socioeconomic status, parental level of education, home possessions, etc. in addition to individuals' total scores may explain performance differences it is necessary to use other matching variables apart from the total score and to purify the total score of the items with DIF in DIF studies. However, DIF studies are conducted by ignoring the aforementioned properties. They are either conducted by using a single

matching variable such as total score, or they are performed based on the total score including the items tagged with DIF. These might be considered sources of errors in DIF studies. Considering all these problems and drawbacks in DIF studies may lead to erroneous implications, the current study employing purified total score and other matching variables apart from the total score was conducted. As a result, this study was required to examine the effect of using other matching variables such as maternal education level, paternal education level and home possessions in addition to the total score in DIF studies and the effect of purified matching variable on DIF determination.

The general purpose of this study is to determine whether the items in the reading literacy test of PISA 2009 display DIF between the samples of Turkey and the USA by using univariate and multivariate matching methods (before and after purifying the total score of the items with DIF). Within this general purpose, answers to the following research questions were sought:

1. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the univariate logistic regression technique before purifying the total score of the items with DIF?
2. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the multivariate logistic regression technique before purifying the total score of the items with DIF?
3. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the univariate logistic regression technique after purifying the total score of the items with DIF?
4. Items in the PISA 2009 reading skills measure display DIF between Turkish and US samples according to the multivariate logistic regression technique after purifying the total score of the items with DIF?

## 2. METHOD

This study, which aims to identify if the items in the PISA 2009 reading skills instrument display DIF between Turkish and US samples by using univariate and multivariate matching methods is a type of correlational survey research design (Tabachnick & Fidell, 2013). Correlational survey design is used to determine the existence of co-variation between two or more variables (Karasar, 2011).

### 2.1. Sample

The population of PISA includes students in the age group of 15 in each participating country. In participating countries, the target population includes all students between the ages of 15 years and 3 months and 16 years and 2 months who are attending school. The sampling strategy of PISA is a two-stage stratified sampling. In the first stage, schools with students in the age group of 15 are selected. In the second stage, students are drawn from the sampled schools (Organisation for Economic Co-operation and Development, 2014). Within the framework of this research, studies were performed on the booklets numbered 1, 3, 4, and 6, in which the OECD has revealed the largest number of items, and the Turkish and US samples who responded to the items in these booklets. The Turkish sample includes 1533 students while the US sample includes 1611 students.

### 2.2. Obtaining Data

The data for this research includes the responses of Turkish and U.S. students to nine items from booklets 1, 3, 4, and 6 of the PISA 2009 reading literacy test, which contained the highest number of items released by the OECD. The data were accessed from the official page of the OECD (http://www.oecd.org/pisa/data/). Six of the nine items in the booklets were selected-response and three were constructed-response. Constructed-response items are dichotomous items that are scored 1-0. For that reason, open-ended items do not have partial scores.

## 2.3. Data Analysis

### 2.3.1. *Testing dimensionality*

It is argued that the multidimensionality of items leads to DIF. For this reason, unidimensionality is a requirement for DIF identification methods that require unidimensionality (Wen, 2014). Confirmatory factor analysis was utilized to test dimensionality and the results are shown in Table 1.

**Table 1.** *Goodness of fit measures estimated from Turkish and US samples.*

| Indices of goodness of fit | Turkish Sample | US Sample |
|---|---|---|
| $\chi^2/df$ | 1.328 | 1.948 |
| CFI | .991 | .987 |
| GFI | .995 | .992 |
| RMSEA | .015 | .024 |

The results estimated based on confirmatory factor analysis support the unidimensionality assumption. In other words, the unidimensional factor model fits the reading literacy data of Turkey excellently, and the USA as seen in Table 1 (Hu & Bentler, 1999; Mcdonald & Ringo Ho, 2002). It could be stated that the factor structure of the reading literacy test is invariant across language groups.

### 2.3.2. *DIF detection technique*

In this study, logistic regression was used as a DIF detection technique. In logistic regression analysis used to determine DIF, variables are included in the model hierarchically. "In Step 1, the matching variable is introduced into the model as an independent variable. In Step 2, the group variable is added. In Step 3, the interaction term is incorporated into the equation. In logistic regression, the chi-square test is used to assess statistical significance, and the contribution of each variable to the model is evaluated. The chi-square value from the first model is then subtracted from the value obtained in the third model. The chi-square value obtained is compared with the chi-square distribution with 2 degrees of freedom. Degrees of freedom 2 is calculated by subtracting the degrees of freedom in the first model (1) from the degrees of freedom in the third model (3) (Crane et al, 2006; Gierl et al, 2000; Hidalgo & Lopez-Pina, 2004; Jodoin, 1999; Sireci & Rios, 2013; Zheng et al., 2007). The result obtained by subtracting the $R^2$ value obtained from the third model from the $R^2$ value obtained from the first model provides evidence for the effect size of DIF (Sireci and Rios, 2013; Zumbo, 1999). Logistic regression can also be applied when more than one variable is used to match individuals (Sireci & Rios, 2013). Nagelkerke $R^2$ value can be employed as an effect size to determine the magnitude of DIF. In order to claim that there is a DIF, the difference in $R^2$ values between models should be at least .13 (Zumbo, 1999). Zumbo and Thomas (1997) suggested the cut-off points in Table 2 for $\Delta R^2 = R^2$ (M3) - $R^2$ (M1) to be used in interpreting the magnitude of DIF for logistic regression (cited in Hidalgo and Lopez-Pina, 2004).

**Table 2.** *Cutt-of points for logistic regression $\Delta R^2$ value.*

| $\Delta R^2$ | *DIF level* |
|---|---|
| $\Delta R^2 < 0.13$ | A level DIF (No DIF or might be neglected). |
| $0.13 \leq \Delta R^2 < 0.26$ | B level DIF (Moderate DIF). |
| $\Delta R^2 \geq 0.26$ | C level DIF (Serious DIF). |

### 2.3.3. *Detection of additional matching variables*

A literature review was conducted to determine matching variables that may be related to reading skills in addition to the total score. Later on, multiple linear regression was carried out to determine the variables of which regression coefficients are significant. The results belonging to multiple linear regression are presented in Table 3.

**Table 3.** *Variables and regression coefficients based on multiple linear regression analysis.*

| Variables | Regression coefficients | |
|---|---|---|
| | β | Standardised Beta |
| Maternal education level | .21 | .17* |
| Paternal education level | .17 | .13* |
| Attitude towards school | .04 | .02 |
| Home possessions | .21 | .10* |
| Family wealth | .03 | .01 |

*$p<0.05$

Table 3 indicates that maternal education level, paternal education level, and home possessions are significant indicators of reading literacy. For this reason, these three variables were considered additional matching variables, alongside the total score on the reading literacy test.

## 3. RESULTS

This section presents the findings obtained in line with the sub-questions of the study. The findings obtained from univariate and multivariate matching-based DIF analyses conducted before and after the purifying the total score of DIF items were compared.

### 3.1. Results Regarding Univariate DIF Before Purification

Table 4 indicates the logistic regression-based univariate DIF analysis performed before purifying the total score. Table 4 indicates that four of the nine items display significant DIF between the Turkish and US samples. The results reveal that all 4 items contain DIF at level A.

**Table 4.** *DIF results based on univariate matching.*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .008* | A |
| R414Q06 | .004* | A |
| R414Q09 | .003 | |
| R414Q11 | .006* | A |
| R452Q03 | .003 | |
| R452Q04 | .001 | |
| R452Q07 | .004* | A |
| R458Q01 | .003 | |
| R458Q07 | .000 | |

*$p<0.05$

### 3.2. Results Regarding Multivariate DIF Before Purification

#### 3.2.1. *Bivariate DIF analysis*

Table 5 indicates the logistic regression-based bivariate DIF analysis performed before purifying the total score. Based on Table 5, four of the nine items displayed significant DIF between the Turkey sample and the US sample. The results reveal that all four items contain DIF at level A. In addition, when compared to univariate DIF analysis, the use of the maternal education level variable apart from the total score did not lead to any change in the number of items labeled as having DIF.

**Table 5.** *DIF results based on bivariate matching (total score plus maternal education level).*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .004* | A |
| R414Q06 | .004* | A |
| R414Q09 | .001 | |
| R414Q11 | .006* | A |
| R452Q03 | .003 | |
| R452Q04 | .002 | |
| R452Q07 | .005* | A |
| R458Q01 | .002 | |
| R458Q07 | .001 | |

*$p<0.05$

### 3.2.2. *Trivariate DIF analysis*

Table 6 indicates the logistic regression-based trivariate DIF analysis performed before purifying the total score.

**Table 6.** *DIF results based on trivariate matching (total score plus maternal education level plus paternal education level).*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .004 | |
| R414Q06 | .004 | |
| R414Q09 | .001 | |
| R414Q11 | .006* | A |
| R452Q03 | .003 | |
| R452Q04 | .003 | |
| R452Q07 | .006* | A |
| R458Q01 | .003 | |
| R458Q07 | .002 | |

*$p<0.05$

Table 6 indicates that two of the nine items show a significant DIF between the Turkish and US samples. The results reveal that both items show level A DIF. Compared to the univariate DIF analyses, the use of the variables of maternal education level and paternal education level in addition to the total score lessened the number of items labeled as DIF from four to two.

### 3.2.3. *Quadrivariate DIF analysis*

Table 7 shows the logistic regression-based quadrivariate DIF analysis performed before purifying the total score.

**Table 7.** *DIF results based on quadrivariate matching (total score plus maternal education level plus paternal education level plus home possessions).*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .004 | |
| R414Q06 | .002 | |
| R414Q09 | .002 | |
| R414Q11 | .006 | |
| R452Q03 | .003 | |
| R452Q04 | .004 | |
| R452Q07 | .006 | |
| R458Q01 | .003 | |
| R458Q07 | .003 | |

According to Table 7, no item displayed DIF between the Turkish and US samples. As a result, compared to univariate DIF analyses, the use of other predictor variables apart from the total score reduced the number of items labeled as DIF from four to zero.

### 3.3. Results Regarding Univariate DIF After Purification

Table 8 indicates the logistic regression-based univariate DIF analysis performed after purifying the total score. According to Table 8, three of the nine items displayed significant DIF between the Turkish sample and the US sample. The results show that all three items contain DIF at level A. It is seen that purifying the total score off the items with DIF reduced the number of items flagged with DIF into three.

**Table 8.** *DIF results based on univariate matching.*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .018* | A |
| R414Q06 | .003 | |
| R414Q09 | .002 | |
| R414Q11 | .005* | A |
| R452Q03 | .002 | |
| R452Q04 | .001 | |
| R452Q07 | .003* | A |
| R458Q01 | .002 | |
| R458Q07 | .001 | |

*$p<0.05$

### 3.4. Results Regarding Multivariate DIF After Purification

#### 3.4.1. *Bivariate DIF analysis*

Table 9 indicates the logistic regression-based bivariate DIF analysis performed after purifying the total score. According to Table 9, three of the nine items displayed significant DIF between the Turkey sample and the US sample. The results demonstrate that all three items contain DIF at level A. Moreover, when compared with the univariate DIF analysis, the number of the items tagged with DIF remained the same.

**Table 9.** *DIF results based on bivariate matching (purified total score plus maternal education level).*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .005* | A |
| R414Q06 | .008* | A |
| R414Q09 | .001 | |
| R414Q11 | .005* | A |
| R452Q03 | .003 | |
| R452Q04 | .001 | |
| R452Q07 | .004 | |
| R458Q01 | .001 | |
| R458Q07 | .002 | |

*$p<0.05$

#### 3.4.2. *Trivariate DIF analysis*

Table 10 indicates the logistic regression-based trivariate DIF analysis performed after purifying the total score. According to Table 10, two of the nine items displayed significant DIF between the Turkish sample and the US. The results reveal that both items contain DIF at level A. Compared to the univariate DIF analysis, the use of maternal education level and

paternal education level variables in addition to the adjusted total score decreased the number of items labelled as DIF from three to two.

**Table 10.** *DIF results based on trivariate matching (purified total score plus maternal education level plus paternal education level).*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .005 | |
| R414Q06 | .007* | A |
| R414Q09 | .002 | |
| R414Q11 | .005* | A |
| R452Q03 | .003 | |
| R452Q04 | .002 | |
| R452Q07 | .004 | |
| R458Q01 | .003 | |
| R458Q07 | .002 | |

*$p<0.05$

### 3.4.3. *Quadrivariate DIF analysis*

Table 11 demonstrates the logistic regression-based quadrivariate DIF analysis performed after purifying the total score.

**Table 11.** *DIF results based on quadrivariate matching (purified total score plus maternal education level plus paternal education level plus home possessions).*

| Item Number | ($\Delta R^2$) | DIF Level |
|---|---|---|
| R414Q02 | .006 | |
| R414Q06 | .004 | |
| R414Q09 | .003 | |
| R414Q11 | .004 | |
| R452Q03 | .003 | |
| R452Q04 | .004 | |
| R452Q07 | .004 | |
| R458Q01 | .003 | |
| R458Q07 | .002 | |

Table 11 shows that none of the nine items were tagged with DIF between the Turkish sample and the US sample. When compared with univariate DIF analyses, it is seen that the use of other predictor variables apart from the purified total score reduced the number of items labeled as DIF from three to zero.

## 4. DISCUSSION and CONCLUSION

It was found that the use of other matching variables apart from the total score led to a decrease in the number of DIF items in general. When the univariate matching method was used, while four items were labeled as having DIF between Turkish and US students using the univariate matching method, none of the items were labeled as having DIF in the DIF analysis based on four-variable matching. Based on this point, it can be argued that additional matching variables explain the DIF displayed by the items in univariate DIF analyses and lead to a reduction in the first type error. This finding is compatible with the findings of studies (Arıkan et al., 2018; Çet, 2006; Roussos & Stout, 1996; Yıldırım & Yıldırım, 2011; Yılmaz, 2021) that examine the effect of using additional matching variables on DIF identification. While some items examined in the study were labeled as DIF in univariate DIF analyses, it was concluded that these items did

not show DIF when additional matching variables were used apart from the total score. Considering that the identified matching variables explain the DIF displayed by these items, it may be recommended to conduct a DIF analysis based on multivariate matching to control the first type of error in DIF studies.

It was determined that carrying away DIF items from the total score caused a variation in the number of items labeled as DIF although it did not yield consistent results. In other words, it can be argued that removing DIF items from the total score does not yield consistent results. This finding is parallel with the findings of studies (French & Maller, 2007; Lee & Geisinger, 2016; Svetina & Rutkowski, 2014). It was revealed that the exclusion of DIF items (removal of DIF items) while calculating the total score, which is the matching variable, affects the DIF detection power of the DIF detection technique. To eliminate the error caused by including DIF items in the total score calculation in DIF studies, and to balance Type I error and test power, it is considered appropriate to exclude DIF items from the total score.

One of the most basic assumptions of international assessment studies is that tests are equivalent in all languages or cultures. However, even in DIF analysis based on multivariate matching, some items were found to have displayed DIF. Considering that the poor quality of the translation makes the validity of the test scores, and therefore the comparability and interpretation of the scores impossible (Gierl, 2000), it is thought that translations in cross-cultural assessment studies should be done with an adaptation approach. However, since the selection of reading texts is of great importance in cross-cultural assessment studies (Grisay, Gonzalez & Monseur, 2009), the selection of these texts can be given particular importance.

In this study, multivariate DIF studies through logistic regression were performed. In a future study, a multivariate DIF analysis could be conducted based on IRT. Additionally, the removal of DIF items from the total score in this study was performed using logistic regression. A similar DIF study could also employ the Mantel-Haenszel method or another suitable DIF detection technique. Furthermore, this study utilized a literature review and multiple linear regression analysis to identify additional matching variables. In future research, alternative statistical methods, such as multilevel modeling, or judgmental approaches could be used to identify extra matching variables.

## Acknowledgments

This study is a part of the doctoral dissertation of the first author under the supervision of the second author.

## 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

**Ahmet Yıldırım:** Literature Review, Resources, Methodology, Data Analysis, Reporting and Writing-original draft. **Nizamettin Koç:** Supervision. Authors may edit this part based on their case.

## Orcid

Ahmet Yıldırım https://orcid.org/0000-0002-0856-9678
Nizamettin Koç https://orcid.org/0000-0001-5412-0727

## REFERENCES

American Educational Research Association. (2014). *Standards for educational and psychological testing*. Washington, DC.

Allalouf, A., Hambleton, R.K., & Sireci, S.G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185-198.

Allalouf, A., & Sireci, S.G. (1998, April). *Detecting sources of DIF in translated verbal items* [Paper presentation]. American Educational Research Association 1998. San-Diego.

Arıkan, S., Van de Vijver, F.J.R., & Kutlay, Y. (2018). Propensity score matching helps to understand sources of DIF and mathematics performance differences of Indonesian, Turkish, Australian, and Dutch students in PISA. *International Journal of Research in Education and Science*, *4*(1), 69-81.

Arffman, I. (2010, August). *Identifying translation-related sources of differential item functioning in international reading literacy assessments* [Paper presentation]. European Conference on Educational Research 2017. Helsinki.

Boughton, K.A., Gierl, M.J., & Khaliq, S.N. (2000, May). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance* [Paper presentation]. Canadian Society for Studies in Education. Alberta.

Braun, M., & Harkness, J.A. (2005). Text and context: Challenges to comparability in survey questions. Zlotnik, J.H.P. & Harkness, J. (Eds.). *Methodological aspects in cross-national research* (pp. 95-107). Mannheim: Zuma.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*(2), 129-147.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning items. *Educational Measurement: Issues and Practice*, *17*, 31-44.

Crane, P.K., Gibbons, L.E., Jolley, L., & Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care*, *44*(11), 115-123.

Çet, S. (2006). *A multivariate analysis in detecting differentially functioning items through the use of programme for international student assessment (PISA) 2003 mathematics literacy items* [Unpublished doctoral dissertation, Orta Doğu Teknik Üniversitesi]. Ankara.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. P.W. Holland & H. Wainer (Eds.). *Differential item functioning* (p. 35-66). New Jersey: Lawrence Erlbaum Publishing.

Ercikan, K., Gierl, M.J., McCreith, T., Gautam, P., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, *17*(3), 301-321.

French, B.F., & Maller, S.J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373-393.

Furlow, C.F., Ross, T.R., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement*, *33*(6), 441-464.

Gierl, M.J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, *25*(4), 280-296.

Gierl, M.J. (2004, April). *Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences* [Paper presentation]. American Educational Research Association. San Diego.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000, April). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large* [Paper presentation]. American Educational Research Association 2000. New Orleans.

Gierl, M.J., & Khaliq, S.N. (2000, April). *Identifying sources of differential item functioning on translated achievement tests: A confirmatory analysis* [Paper presentation]. National Council on Measurement in Education 2000. Louisiana, New Orleans.

Gierl, M.J., Rogers, W.T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF* [Paper presentation]. National Council on Measurement in Education 1999. Montréal, Quebec.

Gradshtein, M.F., Mead, A.D. & Gibby, R.E. (2010). *Making cognitive ability selection tests indifferent across cultures: The role of translation vs. national culture in measurement equivalence*. Retrieved October 20, 2015, from http://mypages.iit.edu/~mead/GradshteinMeadGibby-2010-10-01.pdf

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 63-83.

Hambleton, R.K. (1993). *Translating achievement tests for use in cross-national studies*. Retrieved December 21, 2016, from http://files.eric.ed.gov/fulltext/ED358128.pdf

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. (1993). *Advances in the detection of differentially functioning test items*. Retrieved October 20, 2016, from http://files.eric.ed.gov/fulltext/ED356264.pdf

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), 1-30.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Hidalgo, M.D., & Lopez-Pina, J.A. (2004). Differential item functioning detection and effect size: a comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

International Association for the Evaluation of Educational Achievement. (2012). *TIMSS 2011 international results in mathematics*. Lynch School of Education, Boston College.

Jodoin, M.G. (1999). *Reducing Type I error rates using an effect size measure with the logistic regression procedure for DIF detection* [Unpublished Master's Thesis, University of Alberta]. Alberta.

Karasar, N. (2011). *Bilimsel araştırma yöntemleri* [*Scientific research methods*]. Ankara: Nobel Publishing.

Khalid, M.N., & Glas, C.A.W. (2013). A step-wise method for evaluation of differential item functioning. *Journal of Quantitative Methods*, *8*(2), 25-47.

Lee, H., & Geisinger, K.F. (2016). The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educational and Psychological Measurement*, *76*(1), 141-163.

Mcdonald, R.P., & Ringo Ho, M. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64-82.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2015). *International large-scale assessments: Origins, growth and why countries participate in PISA*. OECD Publishing.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, *6*(2), 1-3.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355-371.

Sireci, S.G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, *20*(2), 148-166.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(2-3), 170-187.

Sireci, S.G., & Swaminathan, H. (1996). Evaluating translation equivalence: So what's the big DIF? Retrieved October 19, 2015, from http://files.eric.ed.gov/fulltext/ED428119.pdf

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. Retrieved October 20, 2015, from http://www.largescaleassessmentsineducation.com/content/pdf/s40536-014-0004-5.pdf

Tabachnick, B., & Fidell, L. (2013). Using multivariate statistics (5th Edition). Allyn & Bacon/Pearson Education.

Van de Vijver, F., & Tanzer, N.K. (2004). *Bias and equivalence in cross-cultural assessment: An overview*. Retrieved October 21, 2015, from http://resilienceresearch.org/files/article-vandevijver_tanzer.pdf

Wen, Y. (2014). *DIF analyses in multilevel data: Identification and effects on ability estimates* [Unpublished doctoral dissertation, University of Wisconsin-Milwaukee]. Wisconsin.

Yıldırım, H.H., & Yıldırım, S. (2011). Correlates of communalities as matching variables in differential item functioning analyses. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 40*, 386-396.

Yılmaz, M. (2021). Eğilim puanları kullanılarak abide çalışmasındaki maddelerin değişen madde fonksiyonu açısından incelenmesi [Unpublished Master's Thesis, Hacettepe University]. Ankara.

Zheng, Y., Gierl, M.J., & Cui, Y. (2007). *Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and logistic regression procedures* [Paper presentation]. National Council on Measurement in Education 2007. Chicago.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistics regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233.

International Journal of Assessment Tools in Education
2024, Vol. 11, No. 4, 787–803
https://doi.org/10.21449/ijate.1529210
journal homepage: https://dergipark.org.tr/en/pub/ijate
*Research Article*

# Examining the use of bilingual accommodations in digital math assessments: User perceptions

**Alexis A. Lopez**[1*]

[1]660 Rosedale Road, Princeton, NJ 08541, U.S.A.

**ARTICLE HISTORY**

*Received: Aug. 07, 2024*

*Accepted: Oct. 15, 2024*

**Keywords:**

Bilingual accommodations,
Digital math assessments,
Multilingual learners,
Linguistic repertoire,
Language modalities.

**Abstract:** Digital math assessments with bilingual accommodations allow multilingual learners to use their entire linguistic repertoire to showcase their knowledge and skills. The bilingual accommodations, which include tools like language translation and audio prompts in both English and Spanish, are designed to be adaptable, giving multilingual learners the freedom to view or listen to the items in either language and to write or say their responses in either language or a combination of both. This study examined how 56 middle school emergent bilingual learners used these bilingual accommodations and explored the perceptions of teachers and students regarding these accommodations. This study provides evidence regarding using bilingual accommodation in math assessments for middle school emergent multilingual learners. The results showed how students used their full linguistic repertoire and language modalities to showcase their math knowledge and skills. Both teachers and students reported having positive perceptions of the bilingual accommodations, reinforcing its responsiveness to different learners' needs and preferences.

## 1. INTRODUCTION

Most current academic content assessments (e.g., math, science) reflect a monolingual view of language and tend to ignore the complex discursive practices used by multilingual speakers (Ascenzi-Moreno et al., 2023; López et al., 2017; Shohamy, 2011). From a monolingual perspective, languages are treated as separate entities and not as a unified system that utilizes the resources of all the languages. Consequently, academic content assessments that reflect a monolingual perspective expect all students to use one language, even if they have multiple languages in their repertoires. However, it is essential to recognize that multilingual learners, when given the opportunity to utilize their entire linguistic repertoire, have the potential to excel in these assessments. Some scholars have pointed out the need to improve existing academic content assessments and develop new ones sensitive to multilingual learners' heterogeneous practices (e.g., García, 2009; López et al., 2017; Otheguy et al., 2019; Sanchez et al., 2013).

Math assessments with bilingual supports utilize the best practices of today's classroom, treating multiple languages as a single, dynamic, unified system. Math assessments conceived in this light allow multilingual learners to utilize their linguistic repertoire more fully by

*CONTACT: Alexis A. LOPEZ alopez@ets.org 660 Rosedale Road, Princeton, NJ 08541, U.S.A.



e-ISSN: 2148-7456

interchangeably moving back and forth from one language to another whenever needed. By doing so, multilingual learners could meaningfully demonstrate their math knowledge and skills during their test-taking experience. It is crucial to note that multilingual learners are unfairly disadvantaged when they are not permitted to draw upon their diverse linguistic repertoire. This is a challenge that needs to be addressed urgently. Consequently, assessments should be designed to value this linguistic diversity and provide multilingual learners with opportunities to demonstrate their knowledge and skills in ways that align with their strengths and preferences and reflect how multilingual learners utilize multiple languages in their daily lives (Ascenzi-Moreno et al., 2023; Paradis et al., 2010).

## 2. BACKGROUND

### 2.1. Accommodations in Content Assessments

To address language challenges in standardized content assessments, educators often use accommodations, which are supports provided during the assessment to help emerging multilingual learners. The goal of accommodations in content assessments is to make the assessment content accessible to all students and is intended to increase the validity of the interpretations of what these learners know and can do in a content area (Abedi, 2014; Roohr & Sireci, 2017). Assessment accommodations can include linguistic modifications, which are changes made to the language of the assessment to make it more understandable for the learner, extended time, and alternative response formats (Abedi et al., 2004; Rios et al., 2020). Typically, schools in the United States (U.S.) offer digital assessments with built-in accommodations for these learners, although accommodations are only available to students with identified needs, as determined by educators based on the student's language proficiency and other factors (Rios et al., 2020).

Despite the widespread use of assessment accommodations for emerging multilingual learners, their effectiveness remains unclear (Rios et al., 2020). It is important to highlight, however, that most of these studies have focused on how accommodations influence changes in test scores rather than their overall impact on accessibility (Li & Suen, 2012; Rios et al., 2020; Wolf et al., 2012). Several meta-analyses have shown only small improvements in test scores (Gezer et al., 2023; Kieffer et al., 2009; Li & Suen, 2012; Pennock-Roman & Rivera, 2011; Rios et al., 2020). Nonetheless, many studies have shown that the most effective accommodations in making the assessment linguistically accessible to emergent multilingual learners are language-based accommodations such as using dictionaries, pop-up glossaries, read-alouds, and native language versions of the assessment (Abedi, 2014). Similarly, a few studies have indicated that digital accommodations show promise and could be a significant part of the future of assessment accommodations (e.g., Roohr & Sireci, 2017; Wolf et al., 2021). Some accommodations, such as simplifying language or using glossaries, have shown positive results (Abedi & Lord, 2001; Pennock-Roman & Rivera, 2011). However, other accommodations, like dual language testing or translation, have produced inconsistent findings or lack sufficient research (Pennock-Roman & Rivera, 2011; Rios et al., 2020). This highlights the urgent need for more in-depth research on assessment accommodations to ensure the best outcomes for emerging multilingual learners.

There is growing support for individualized, research-based accommodations and improved teaching methods to help emerging multilingual learners succeed in content assessments (Koran & Kopriva, 2017; Roschmann et al., 2021). Among the challenges that exist in using assessment accommodations include proper implementation (Abedi et al., 2004; Wolf et al., 2012) and the need for tailored approaches based on individual student needs (Bartlett, 2021). To improve assessment validity, researchers recommend developing accommodations that consider students' linguistic needs (Liu, 2023), examining the impact of score interpretation on assessments with accommodations (Iliescu & Greiff, 2022), and investigating the role of academic language skills in content assessments (Kieffer et al., 2009). This research is crucial

in ensuring the best outcomes for emerging multilingual learners and underscores the value of continued research on accommodations in content assessments. The potential of individualized, research-based accommodations is promising and offers hope for improving content assessments for emergent multilingual learners.

### 2.2. Bilingual Assessment Accommodations

Bilingual accommodations in math assessments can support emergent multilingual learners by allowing students to engage in translanguaging (Lopez et al., 2017). Translanguaging refers to "the deployment of a speaker's full linguistic repertoire without regard for watchful adherence to the socially and politically defined boundaries of named languages" (Otheguy et al., 2015, p. 283). Here, 'named languages' refer to social categories such as English or Spanish (Otheguy et al., 2015). However, the named languages are presented separately when using bilingual supports on a digital math assessment (Lopez et al., 2017).

As a result, a digital math assessment with bilingual accommodations can be seen as an assessment that empowers emergent multilingual learners to utilize their entire linguistic repertoire and language modes to showcase their math knowledge and skills (Lopez et al., 2017). Their linguistic repertoire encompasses standard and vernacular language varieties (Sayer, 2013). The goal is to foster linguistically adaptive bilingual practices within a single assessment context (Shohamy, 2011) and allow students to utilize different semiotic resources, enabling them to perform in writing or orally (Li, 2011) to demonstrate what they know and can do. The items are in multiple languages (e.g., English and Spanish). However, it is the students who have the autonomy to select the named language and the language mode they prefer to use to demonstrate their math knowledge and skills (Lopez et al., 2017).

Several bilingual accommodations have been documented to effectively reduce the score gap between emergent multilingual learners and non-multilingual learners attributed to emergent multilingual learners' limited proficiency in English (Francis et al., 2006). Bilingual accommodations include bilingual test forms, pop-up bilingual glossaries, reading aloud the directions and items in English and the home language, and allowing students to respond in the home language (Abedi, 2009; Pennock-Roman & Rivera, 2011). Although test translation is commonly used as an accommodation to support emergent multilingual learners, not all of them may benefit from this type of support because their language and literacy proficiencies in English and their home language vary tremendously (Smarter Balanced Assessment Consortium, 2012; Solano-Flores, 2008). Thus, it is vital for educators and policymakers to provide bilingual accommodations that meet the specific needs of emergent multilingual learners (Koran & Kopriva, 2017). To enable the agency of emergent multilingual learners and empower them to select which bilingual accommodation they want or need to use, these accommodations should always be available to the students (Lopez et al., 2017).

The evidence on the impact of bilingual accommodations in reducing the achievement gap between multilingual learners with emergent English skills and native English speakers is inconclusive. However, there is support for using bilingual accommodations to make content assessments more equitable and unbiased for multilingual learners (Goodrich et al., 2021; López et al., 2015). Bilingual accommodations have also been found to be effective in helping multilingual learners access the content of assessment items (Abedi, 2021; Roschmann et al., 2021). Therefore, it is important to continue providing empirical evidence that bilingual accommodations do not threaten the validity of content assessments and make them accessible for multilingual learners.

## 3. METHOD

### 3.1. The Purpose of the Study

I used a concurrent mixed methods approach where quantitative and qualitative data were combined to examine the use of bilingual accommodations on digital content assessments (e.g.,

math, science). To focalize the study, I selected a digital math assessment to measure the math knowledge of middle school multilingual learners with emerging English skills. I examined which accommodations the students used and how often they used them and investigated teachers' and students' perceptions about using bilingual accommodations. The findings of this study can be directly applied to improve the learning experience of these students, making the research highly relevant and helpful. The following highly relevant research questions guided this study:

1. How did emergent multilingual learners use the bilingual accommodations on a digital math assessment?
2. What perceptions did emergent multilingual learners have of the bilingual accommodations' usefulness in measuring their math knowledge?
3. What perceptions did middle school math teachers have of the bilingual accommodations' usefulness in measuring students' math knowledge?

### 3.2. The Digital Math Assessment

The digital math assessment used in this study was developed for research purposes only, and the performance on the assessment did not impact the student's grades or standing in their math classes. The assessment aimed to measure students' knowledge of ratios and proportional relationships as described by the U.S. Grade 6 Common Core State Standards for Mathematical Practice (CCSSO, 2010). The digital math assessment was developed using an evidence-centered design (ECD) framework (Mislevy et al., 2003) to ensure its validity from the outset (Kobrin, 2022). Moreover, two math teachers independently reviewed all the items to evaluate the relevance and representativeness of the content domain. The two math teachers also provided suggestions for improving the items, ensuring the quality of the digital math assessment tool. The items were first developed in English and then translated into Spanish. Two bilingual math teachers reviewed the translated items to evaluate the quality and accuracy of the translated items and to ensure both language versions measured the same construct at the same difficulty level.

The math assessment was delivered on a digital platform and contained nine items with bilingual accommodations, including 13 multiple-choice questions and three constructed-response questions. The constructed-response questions had two parts. Part A included number entry questions and Part B included a constructed-response question. This student-centered approach ensured that the assessment was designed with the best interests of the students in mind, allowing them to demonstrate their knowledge in the most effective way. Of a possible score of 19, the scores of all 56 participants ranged from 2 to 15, with a mean of 4.7. The standard deviation was $SD = 3.02$. Cronbach's alpha reliability estimate for the multiple-choice questions was .81, indicating fair consistency of measurement across individual items. The inter-rater reliability of the scoring of the constructed response questions was high, as indicated by an exact agreement of 94% and a Kappa index of 87%. The standard error of measurement was .403.

To allow the students to use their entire linguistic repertoire and language modes, several comprehensive bilingual accommodations were added. These accommodations were always available so students could use them at any given time, if needed. Initially, the students saw the items in English, but they could also see them in Spanish by clicking on a button; they could also toggle back and forth between language tabs at any time (bilingual accommodation 1). For constructed-response questions, students could write their responses in either language, using any dialect, or a combination of both (bilingual accommodation 2). Alternatively, students could also record their responses in either language or a combination of both (bilingual accommodation 3). A few non-mathematical-related words were highlighted in the English or Spanish tab. If students clicked on the highlighted words, they saw a pop-up glossary with synonyms for these words to account for dialect variation (bilingual accommodation 4). This

support did not apply to math-related terminology, which was construct-relevant. The words in the English version were selected based on how critical they were to understand the question. The words in the Spanish version were selected based on how different they were in terms of variety or region. For example, the word "plátano" is also known as "banana," "banano," "cambur" and "guineo" in different Spanish varieties or different regions in Latin America. Thus, we highlighted the word "plátano" in Item 8 and added the other expressions in the pop-up glossary. Finally, students could click on an avatar's picture to listen to someone read aloud the directions and the questions in English and Spanish, depending on the language tab they select (bilingual accommodation 5). This comprehensive approach ensures that the assessment is inclusive and supportive of all students, regardless of their linguistic background.

### 3.3. Participants

For this study, I selected schools using a combination of purposive and convenience sampling. I specifically focused on recruiting schools with a large number of Spanish-English bilingual students. The schools were selected from a pool of institutions that had participated in previous studies. I chose two institutions because they were willing to participate and easily accessible. The use of purposive and convenience sampling was suitable for this exploratory research study, as it helped me gather initial insights about how multilingual learners used bilingual accommodations to complete the items. The study was conducted with 56 students from two schools in two U.S. states, Oregon and Texas – 28 from each state: 11 sixth graders, 36 seventh graders, and 9 eighth graders. The sample was evenly divided between males and females (30 male students, 53.6%). Their age ranged between 11 and 14 years of age, and they spoke English (3 students, 5.4%), Spanish (24 students, 42.9%), or both languages (29 students, 51.8%) at home. Most students (39 students, 69.6%) were born in the U.S. and began attending school in the U.S. either in pre-kindergarten (22 students, 39.3%) or kindergarten (17 students, 30.4%). Of the students who reported being born outside of the U.S., all but one reported being born in Mexico. The teachers rated most of the students' math knowledge as low (40 low, 15 average, 0 high). The students' levels of English language proficiency varied, though all the students were categorized as English learners by their teachers (22 low, 11 average, 23 high). Additionally, 13 middle school math teachers were recruited for a focus group interview. Two focus group interview meetings were scheduled, one with seven teachers (four female teachers, three male teachers) and the other with six (four female teachers, two male teachers). The teachers included in this study met the following criteria: 1) had at least five years of experience teaching emergent multilingual learners, and 2) had at least ten emergent Spanish-speaking bilingual learners in any of their math classes. These teachers were recruited from a pool of teachers who had participated in previous studies in the last five years and were willing to participate in the study.

### 3.4. Procedures

A week before the digital math assessment with bilingual accommodations, teachers were tasked with completing a student background questionnaire. This comprehensive tool was specifically designed to gather detailed information about the participants, such as their age, gender, grade, length of time in the United States, languages spoken at home, and scores on state English language proficiency and math assessments. Additionally, teachers were asked to rate their students' English language and math skills as high, average, and low. These ratings were crucial, as they were based on the students' scores on the annual state-wide English language proficiency summative assessment taken the previous school year, or for new students, on their scores on the initial English language identification assessment taken at the beginning of the current school year. The teachers' judgments on their students' math abilities were based on the students' grades in their math class.

Prior to the digital math assessment, students were actively involved in the process. They filled out an online background questionnaire, providing additional information about their language

and educational background. Then, they took the assessment with bilingual accommodations. The assessment platform automatically recorded their responses, the time they spent on each item, and the number of times they used dual language supports, ensuring the data's accuracy and reliability.

Next, students completed an online questionnaire at the end of the study to gather feedback on their perceptions of the items with dual language. The survey included 10 questions. The first five questions used a 3-point Likert scale ranging from 1 = did not like to 3 = liked a lot to measure how much they liked each bilingual support. The last five questions used a 3-point Likert scale ranging from 1 = not useful to 3 = very useful to rate the perceived usefulness of each bilingual support. Finally, to gather more in-depth insights, two focus group interview meetings with the math teachers were conducted. These meetings were significant as they provided a platform for the teachers to share their experiences in supporting emergent multilingual learners in classroom assessments and to discuss their perceptions of each of the bilingual accommodations. Two focus group interview meetings were scheduled, one with seven teachers and the other with six. Each meeting was audio-recorded and lasted approximately 90 minutes.

### 3.5. Data Analysis

Each student log file (the file generated with each click the student made) was analyzed to determine how the student used the accommodations (research question 1). Frequencies of the times students used each accommodation in English and Spanish were calculated. The students' surveys were analyzed by calculating the frequencies of the ratings on perceptions and usefulness (research question 2). Finally, two researchers worked independently to understand how the math teachers perceived the bilingual accommodations (research question 3), carefully analyzing the two focus group transcripts by identifying key themes and patterns in the participants' responses. This analysis involved multiple rounds of coding and review by two researchers to ensure reliability. The researchers carefully categorized interview sections based on their content (e.g., current practices, perceptions, and recommendations) and then closely examined these categories to find recurring themes (e.g., implementing accommodations in the classroom, usefulness of accommodations, challenges in implementing accommodations). The researchers compared their findings to ensure consistency and resolved disagreements through open discussion. Ultimately, the two researchers identified critical themes related to how teachers currently use accommodations in their classrooms, what they like about the accommodations, what they do not like about them, and other ways to support multilingual learners. The resulting themes revealed how educators used bilingual supports in their classrooms and how helpful each bilingual accommodation in the digital assessment was.

## 4. FINDINGS

### 4.1. Use of Bilingual Accommodations (Research Question 1)

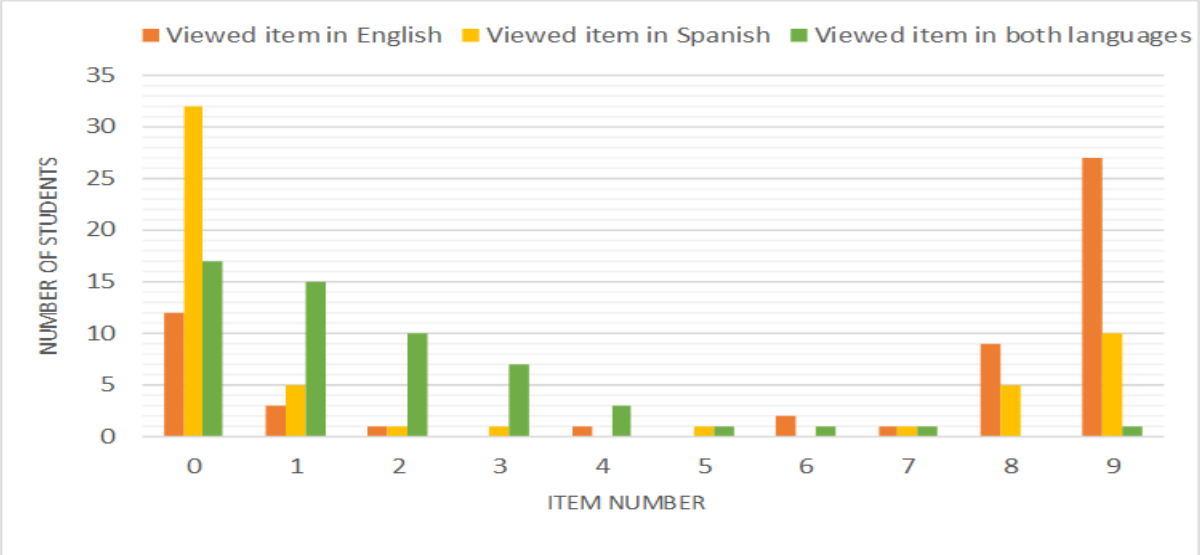
Overall, the students made comprehensive use of the bilingual accommodations, utilizing them frequently. It is noteworthy that all the students made use of the available accommodations at least once, with eleven students using all the available options. Eight students exclusively used the accommodations in English, while six students opted for the Spanish-only accommodations. A significant number of forty-two students utilized the accommodations in both languages. In the following sections, I provide a detailed breakdown of how students utilized each available accommodation.

#### 4.1.1. *Language use*

Students actively participated in the study, using both English and Spanish to answer the items. While most completed the items in English (see Figure 1), it's important to note that 44 students answered at least one of the items in English, with 32 answering all in English. On average, these students answered 7.9 items in English. Conversely, 24 students answered at least one of

the items in Spanish, with 12 answering all in Spanish. On average, 6.5 items were answered in Spanish. Twenty-nine students demonstrated their active involvement by toggling back and forth between English and Spanish in at least one item. One student showed exceptional engagement by switching languages in all the items. In total, 12 students answered items in both languages. Seven answered most of the items in English, with an average of 6.2 items answered in English. Contrarily, five students answered most of the items in Spanish, with an average of 2.8 items answered in Spanish.

**Figure 1.** *Frequency graphs of language used to answer the items.*




When it comes to the language used in the constructed-response items, it is worth noting the efforts of most students to respond in English, even when their proficiency was low (see Table 1). Thirty-three students responded to all the constructed-response items in English, 13 only in Spanish, and two using only symbols and numbers (e.g., math sentences). Three students showcased their individuality through their responses. Two of them answered some questions in English, and some used symbols and numbers. One student answered two questions in English and one in Spanish. In general, students did not mix languages in their responses, with only one student doing so for one question.

**Table 1.** *Type of response in each constructed-response item.*

| Type of response | Q7 | Q8 | Q9 | Total |
|---|---|---|---|---|
| In English | 31 | 32 | 30 | 93 |
| In Spanish | 12 | 10 | 11 | 33 |
| In both English and Spanish | 0 | 1 | 0 | 1 |
| Only symbols and numbers | 3 | 3 | 2 | 8 |
| No response | 10 | 10 | 13 | 33 |

Some students demonstrated resourcefulness in their use of translingual practices when responding to the constructed-response items in English. Translingual practices refer to the "ability to merge different language resources in situated interactions for new meaning construction" (Canagarajah, 2013, pp. 1–2). A few students wrote responses such as, "i oli ad them all up" [I only added them all up], "I ONLI POT 4 BOES" [I only put four boxes] and "i nhou because i didet on mi paper" [I know because I did it on my paper].

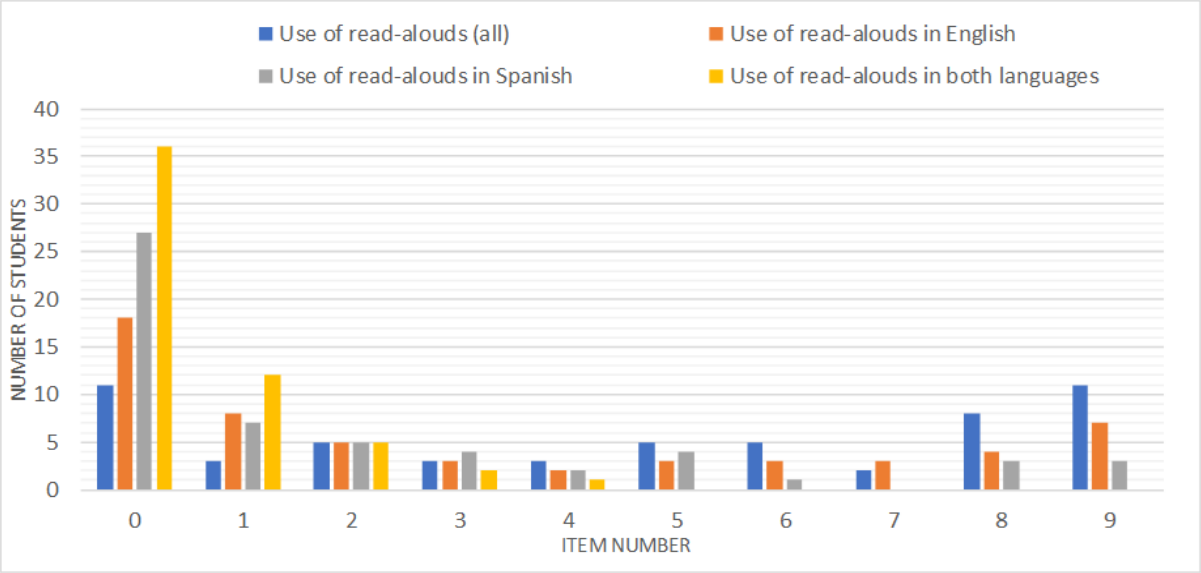
### 4.1.2. *Language modality in constructed-response items*

When it comes to the language mode used to answer the three constructed-response questions, it's worth noting that most participants (except for three students) preferred to type their responses rather than record them. These three students recorded all their responses in English. However, it is important to note that 14 students initially attempted to record their responses but found it challenging and switched to providing a written response. This adaptability suggests that for some, writing their responses was easier than recording them. The majority's preference for typing may indicate a higher comfort level in typing responses over recording them.

### 4.1.3. *Read alouds*

Figure 2 provides information about the number of times students listened to someone reading aloud the questions to them. Students used the read-aloud accommodation frequently; 46 students used it at least once; 11 students used it in all the items. On average, students used the read-aloud accommodation in six items. The read-alouds were used more frequently in English; 38 students listened to the items in English at least once, while 27 students did the same in Spanish. Altogether, students used the read-aloud in English for five of the nine items, while the read-aloud in Spanish was used for four. Notably, 20 students (35.7%) listened to at least one of the items in both English and Spanish, which was an unexpected finding. Overall, these students listened to at least one item in English and Spanish.

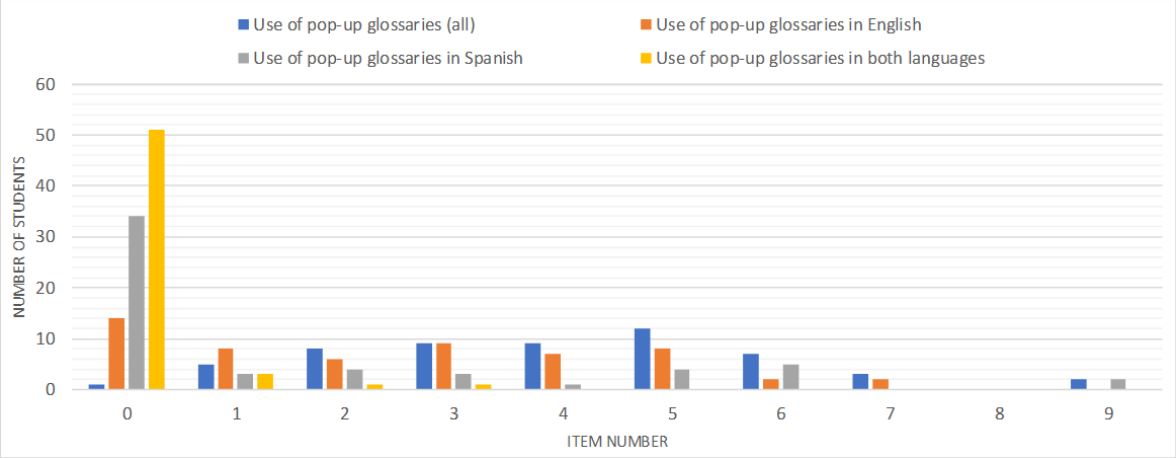
**Figure 2.** *Number of items in which students listened to the questions by language.*




### 4.1.4. *Pop-up glossaries*

Students used the pop-up glossaries frequently (see Figure 3). Only one student did not use this accommodation. On average, students used the pop-up glossaries in four items. When categorized by language, I found that students used the pop-up glossary more frequently in English. Forty-two students used it at least once in English; two of them used this accommodation in seven items. Conversely, 22 students used this accommodation at least once in Spanish; two used it in all the items. Only five students used this accommodation in English and Spanish at least once, demonstrating their adaptability and diverse usage patterns.

**Figure 3.** *Number of items in which students used the pop-up glossaries (by language).*




## 4.2. Students' Perceptions (Research question 2)

When asked to share their thoughts, the students expressed their appreciation for the five bilingual accommodations. Their feedback revealed a general liking for all the available options (see Figure 4). They particularly enjoyed the translations and the read-alouds. Even the least favored accommodations, such as recording responses in constructed-response items and the pop-up glossaries, were still helpful (see Figure 5). The students' appreciation for these accommodations was further confirmed when they reported that they all clarified what the questions were asking them to do and were very helpful when answering them. According to their feedback, the most beneficial bilingual accommodation is viewing the items in both English and Spanish (translation accommodation).

**Figure 4.** *Students' perceptions of the bilingual accommodations (%).*

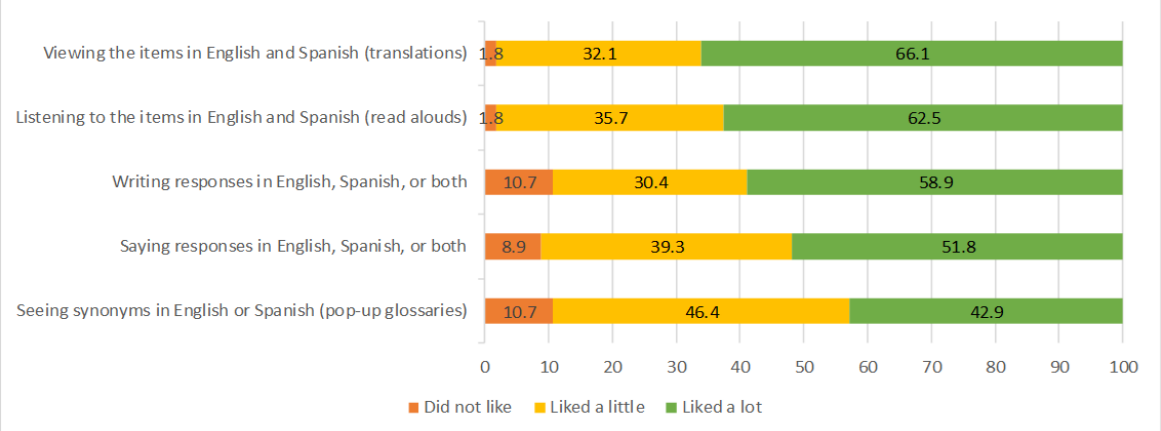
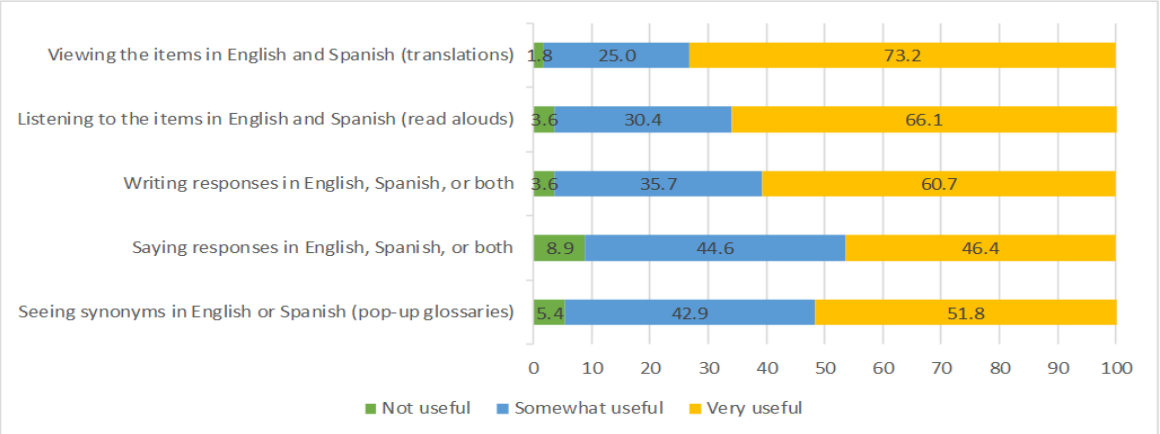


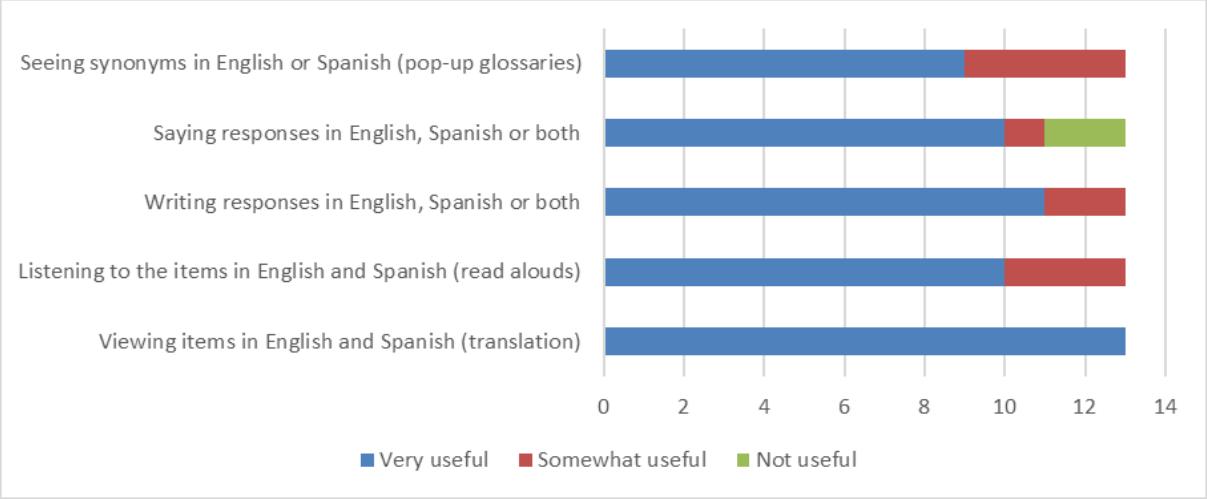

**Figure 5.** *Students' perceptions of the usefulness of the bilingual accommodations (%).*



### 4.3. Teachers' Perceptions (Research Question 3)

In general, the math teachers found all the bilingual accommodations beneficial. They felt that these accommodations are similar to how they support emergent multilingual learners in their classrooms, allowing students to validly demonstrate their math knowledge and skills. Figure 6 indicates the number of teachers who found each bilingual accommodation useful. One teacher expressed this alignment: "I give my assessments in both English and Spanish. Sometimes I use online translators and sometimes I use fellow teachers who, you know help me translate the questions. But I give it in both English and Spanish" (excerpt focus group 1, female teacher 1).

**Figure 6.** *Teachers' perceptions of the usefulness of the bilingual accommodations (N = 13).*




Five teachers commented that they usually have students help each other by translating the items, reading aloud the items, or explaining what the items are asking them to do. One of the teachers explained how they have students help each other in classroom assessments: "I'll find another student who speaks Spanish to explain it. I know sometimes it's difficult, because you worry they might help them with the math process. So, I tell them, if you're helping, you can't tell them what to do. Just explain it so they can understand" (excerpt focus group 2, male teacher 2).

Three teachers did not find the '"Say the Response"' accommodation useful because most of their open-ended questions require students to write math expressions or graph the response. They also mentioned that this accommodation is rarely offered in large-scale state math assessments, so they want their students to become comfortable writing their responses. However, the other ten teachers liked this accommodation. One of them stated the following about allowing students to provide oral responses: '"I think it's important to see how well they can explain it, whether in English or Spanish, writing or speaking. Some of my students can't speak English well and can't read or write in Spanish. So, using this accommodation is the only way they can complete the questions"' (excerpt focus group 2, male teacher 1). It's important to note that while this accommodation may not be suitable for all types of questions, it can significantly benefit students who struggle with written expression, allowing them to demonstrate their understanding of mathematical concepts more effectively.

Moreover, the teachers liked the pop-up glossaries but wanted to change how they were implemented. For example, one of the teachers commented, "If you highlight certain words, you're drawing attention to it. Also, they might stumble across other words they don't know, but they're not highlighted. My students raise their hands in class to ask me about tricky words. I like having these teachable moments in class" (excerpt focus group 1, female teacher 2). Another teacher suggested having pop-up glossaries for all the words because "it is difficult to determine which words are problematic for English learners" (excerpt focus group 1, female teacher 2).

The math teachers were also asked to judge whether adding bilingual accommodations would change the construct measured in the digital assessment. All the teachers agreed that the bilingual accommodations do not change the items' construct. Regarding the construct of the item, one teacher explained: "In the question where they have to find the area of a circle, are we assessing something different if they do it in any language or if they read, listen to, write, or say it? They are still finding the area of the circle. It's assessing the same math" (excerpt focus group 1, male teacher 1).

The teachers also highlighted the need for assessments that help emergent multilingual learners overcome language barriers and demonstrate their math knowledge. They all feel an assessment with bilingual accommodations would be instrumental in learning what newly enrolled students know and can do in math and that they would like more accommodations added. For example, four teachers commented that students can use online translators to translate their responses into English if they respond in their home language. Other suggestions included modifying the language of the items to reduce the reading load (9 teachers) and having more visual representations like graphs or number lines (3 teachers).

## 5. DISCUSSION

First, I delved into the students' utilization of the available bilingual accommodations. As in other studies, I discovered that students employed these accommodations in diverse ways (e.g., López, 2023; López et al., 2019). Notably, some students responded to all the items in a single language (either English or Spanish), while a few seamlessly switched between the two to answer the items. This finding is particularly striking as it demonstrates a high level of bilingual proficiency and the ability to switch languages as per the task at hand. Furthermore, some students opted to use all or some of the available bilingual accommodations, while others chose not to use them at all. Most students utilized the accommodations more frequently when tackling the items in English, indicating their strategic use of the available resources based on their needs. These results echo other studies that show how multilingual learners strategically employ their linguistic resources (e.g., Velasco & García, 2014).

This study also brought to light the creative use of language by some students with emerging English language skills who chose to complete the assessment in English. These students, armed with their emergent English writing skills, tackled the constructed-response questions in a unique way. They deviated from standard written English conventions and produced hybrid responses that incorporated elements from both English and Spanish. In their responses, they used phonemic and phonological features in Spanish to spell some words in English (e.g., “pot” instead of put, “nhou” instead of know, “didet” instead of “did it”). This flexible and inventive language use across linguistic boundaries, often referred to as translingual practices (Canagarajah, 2013), not only underscores their creativity but also their ability to use all their linguistic resources, even if they are not fully developed (Martin-Beltrán, 2014). The concept of translingual practice challenges traditional notions of language boundaries or language separation and emphasizes the fluid, dynamic nature of communication across linguistic and cultural contexts (Canagarajah, 2018).

These findings underscore the importance of expanding scoring to account for translingual responses in academic content assessments for emerging multilingual learners. This approach involves scoring responses regardless of the language or mode used, including mixing or hybridizing the languages. Allowing students to use all their linguistic resources in math assessments, including the use of multiple languages and different modalities (Kusters et al., 2017; Li, 2011), is crucial. In this study, the bilingual accommodations enabled students to use different modalities to interact with and respond to the items. A few students listened to the directions, some in English and some in Spanish. Similarly, a few students also listened to some of the questions, in English or Spanish.

To answer the constructed-response questions, most of the students typed their responses; however, three students used the Say the Response accommodation to record their responses. This diverse use of different bilingual accommodations by the students demonstrates their determination to understand and respond to the items, and to draw on new and complex language practices (García & Li, 2014). Lastly, students used the pop-up glossaries more frequently when viewing the items in English. This suggests that many students preferred to answer the items in English, so it is imperative to provide more accommodations in English. For example, language simplification (e.g., Rivera & Stansfield, 2004), pictorial glossaries (e.g., Turkan et al., 2019), word boxes (e.g., Harmon et al., 2013), or sentence starters/frames (e.g., Donnelly & Roe, 2010).

Second, I also examined how students perceived the available bilingual accommodations. The students positively perceived the bilingual accommodations, even if they did not use them or felt unnecessary. Students liked having this flexibility because the bilingual accommodations are always available and can be used whenever needed. Having assessment accommodations that are always accessible gives emerging multilingual learners 'student agency' (Adie et al., 2018; Emirbayer & Mische, 1998) and enables student choices and actions in digital math assessments. In a way, students are empowered and are more engaged in the assessment because now they have the autonomy to decide if they want to use the bilingual accommodations, when to use them, or which ones to use. One of the main benefits of having increased student agency in assessment includes enhanced student motivation and engagement, which results in students having a more active role in assessment decisions (King et al., 2024).

When it comes to the students' preferences, it was discovered that they favored all the bilingual supports. However, the most popular ones were viewing the items in both English and Spanish and having someone read them aloud. In terms of usefulness, a significant majority of students believed that the bilingual accommodations were instrumental in their understanding and completion of the items. This finding aligns with other studies that have examined the use of assessment accommodations (e.g., López et al., 2019; Wolf et al., 2021). This is a crucial point to highlight, as the primary aim of bilingual accommodations is to enhance the accessibility of the items for emergent multilingual learners (Kieffer et al., 2009; Rios et al., 2020; Wolf et al., 2012). According to the students, these accommodations effectively reduced language barriers, enabling them to better showcase their true math proficiencies.

Finally, this study aimed to explore math teachers' perspectives on the effectiveness of bilingual supports on a digital math assessment. The teachers, in general, found these supports to be beneficial and in line with the supports they offer in their classrooms. This alignment underscores the importance of providing assessment accommodations that students are accustomed to (Rios et al., 2020). Therefore, it is crucial to extend similar supports to emergent multilingual learners in classroom instruction to aid in their academic success.

It is worth noting that teachers discussed the need for assessment accommodations that are tailored to the needs of multilingual learners. Recent research underscores the significance of 'linguistically responsive assessments' for multilingual and diverse learners (e.g., Walker et al., 2023; Yang, 2024). These assessments integrate learners' linguistic and cultural resources, thereby supporting both content learning and language development (Lyon, 2023). 'Linguistically responsive assessments' are those that consider the diverse linguistic backgrounds of students and provide appropriate accommodations to support their learning. A few studies have indicated that multilingual learners perform better on multilingual tasks than on monolingual tasks (e.g., Ascenzi-Moreno, 2018).

### 5.1. Limitations of the Study

The study was largely limited by the fact that the assessment itself was exploratory, and students' performance had no consequences, making it a no-stakes assessment. This could have affected the students' motivation to perform well, which is associated with lower performance

(Wise & DeMars, 2005). Also, there were some limitations with the available student sample, which was homogeneous. The sample did not vary in mathematical proficiency (e.g., most students exhibited low mathematical proficiency), language background, and home demographics. The lack of variance in mathematical proficiency is a significant limitation that prevented the study from adequately exploring the relationship between performance and the use of bilingual accommodations. Despite these limitations, the study's findings on the use of bilingual supports on digital math assessments for middle school multilingual learners with emergent English skills are valuable and of great importance to the field of bilingual education.

### 5.2. Implications for Future Research and Practice

There is a pressing need for further research to validate the use of bilingual accommodations on digital math assessments. This study found that students had a positive perception of all available bilingual accommodations. However, it would be intriguing to investigate if there is a relationship between individual student use of each bilingual accommodation and their preferences. There may be patterns in how students use the accommodations and their preferences, based on student characteristics, educational experiences, or item characteristics, that we have yet to explore. For instance, students may be more likely to use specific bilingual accommodations when faced with assessment items with high language complexity. Follow-up studies could examine students' rationale for using specific accommodations using think-aloud protocols, to understand the reasons behind their bilingual accommodation choices.

These studies can also focus on how specific subgroups of multilingual learners use assessment accommodations, such as students who have learned math mostly in English versus those who have learned math mostly in Spanish. Moreover, future studies should also investigate the innovative potential of leveraging artificial intelligence (AI) to personalize assessment accommodations. The use of AI could improve the way we meet the needs of multilingual learners. By determining the needs based on the characteristics of the students or their educational experiences, we can ensure a more tailored, effective, and inclusive approach to assessment accommodations.

## 6. CONCLUSION

This study provides evidence regarding the use of bilingual accommodation in math assessments for middle school emergent multilingual learners. The students used their full linguistic repertoire to showcase their math knowledge and skills. The bilingual accommodations allowed students to select which language (English, Spanish, or both) they wanted to use to access and understand the items. The bilingual accommodations also allowed the students to select which language they wanted to use to respond to the items and allowed them to use their entire linguistic repertoire to answer the constructed-response questions. In the constructed-response questions, students used English, Spanish, numbers, symbols, or a combination of all these resources to solve the problems and to demonstrate their understanding without being penalized. A few students even used translingual practices to respond to the constructed-response questions. The bilingual accommodations also allowed students to use different language modalities to understand the questions (i.e., view and listen to items in both languages) and to answer the open-ended questions (i.e., say or write their response). This study takes an important first step toward understanding the potential benefits of making use of students' multilingual repertoire in a math assessment. Finally, students and teachers had positive perceptions of the bilingual accommodations and liked that they reduced the language barriers and allowed students to use all their language resources to showcase their math knowledge and skills. Although this study is built around the students' interactions on a particular set of items, the issues raised are likely to be of relevance to other mathematic assessments or other content areas (e.g., science). However, the most significant aspect of this study is its global implications. The prevalence of multilingualism worldwide due to globalization, mobility, and technology (Cenoz & Gorter, 2015) makes the findings from this

study not only relevant but also important for many contexts around the world, underscoring the significance and relevance of the study.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author. **Ethics Committee Number**: The Committee for Prior Review of Research, IRB-FY2023-16.

## Orcid

Alexis Lopez https://orcid.org/0000-0002-4616-1091

## REFERENCES

Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment, 14*(3-4), 195-211.https://doi.org/10.1080/08957347.2014.944310

Abedi, J. (2014). The use of computer technology in designing appropriate test accommodations for English language learners. *Applied Measurement in Education, 27*(4), 261–272. https://doi.org/10.1080/08957347.2014.944310

Abedi, J. (2021). Accommodations and universal design. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed.) (pp. 306–321). Routledge.

Abedi, J., Hofstetter, C.H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1–28. https://doi.org/10.3102/00346543074001001

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Adie, L.E., Willis, J., & Van der Kleij, F.M. (2018). Diverse perspectives on student agency in classroom assessment. *Australian Educational Researcher, 45*, 1-12. https://doi.org/10.1007/s13384-018-0262-2

Ascenzi-Moreno, L. (2018). Translanguaging and responsive assessment adaptations: Emergent bilingual readers through the lens of possibility. *Language Arts, 95*(6), 355–369. https://doi.org/10.1007/s13384-018-0262-2

Ascenzi-Moreno, L., García, O., & López, A.A. (2023). Latinx bilingual students' translanguaging and assessment: A unitary approach. In S. Melo-Pfeifer & C. Ollivier (Eds.). *Assessing plurilingual competence and plurilingual students: Theories, educative issues and empirical approaches* (pp. 48–61). Routledge. (Routledge Research in Language Education) https://doi.org/10.4324/9781003177197

Bartlett, H.J. (2021). Assessments and accommodations for English language learners: A literature review. *The Nebraska Educator: A Student-Led Journal, 60*. https://doi.org/10.32873/unl.dc.ne025

Canagarajah, S. (2013). *Literacy as translingual practice between communities and classrooms.* Routledge.

Canagarajah, S. (2018). Translingual practice as spatial repertoires: Expanding the paradigm beyond structuralist orientations. *Applied Linguistics, 39*(1), 31-54. https://doi.org/10.32873/unl.dc.ne025

Cenoz J, Gorter D. (Eds.). (2015). *Multilingual education: Between language learning and translanguaging*. Cambridge University Press. https://doi.org/10.1017/9781009024655

Council of Chief State School Officers (CCSSO). (2010). Common Core State Standards for Mathematics. CCSSO. Retrieved from https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf

Donnelly, W.B., & Roe, C.J. (2010). Using sentence frames to develop academic vocabulary for English learners. *The Reading Teacher, 64*(2), 131-136. https://doi.org/10.1598/RT.64.2.5

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology, 103*, 962–1023. https://doi.org/10.1086/231294

Francis, D., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions*. RMC Research Corporation, Center on Instruction.

García, O. (2009). *Bilingual education in the 21st century: A global perspective*. Wiley-Blackwell.

García, O., & Li, W. (2014). *Translanguaging: Language, bilingualism and education*. Palgrave Macmillan Pivot.

Gezer, T., Flowers, C., & Lambert, R. (2023). Effectiveness of English learners computer-based testing accommodations: A meta-analysis. *Journal of English Learner Education, 15*(1). https://stars.library.ucf.edu/jele/vol15/iss1/2

Goodrich, J.M., Koziol, N.A., & Yoon, H. (2021). Are translated mathematics items a valid accommodation for dual language learners? Evidence from ECLS-K. *Early Childhood Research Quarterly, 57*, 89–101. https://doi.org/10.1016/j.ecresq.2021.06.001

Harmon, J.M., Fraga, L.M., Martin, E., & Wood, K.D. (2013). Revitalizing word walls for high school English learners: Conventional and digital opportunities for learning new words. *Georgia Journal of Literacy, 36*(1), 20–28. https://doi.org/10.56887/galiteracy.37

Iliescu, D., & Greiff, S. (2022). Some thoughts and considerations on accommodations in testing. *European Journal of Psychological Assessment, 38*(4), 239-242. https://doi.org/10.1027/1015-5759/a000732

Kieffer, M.J., Lesaux, N.K., Rivera, M., & Francis, D.J. (2009). Accommodations for English language learners taking large-scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research, 79*(3), 1168-1201. https://doi.org/10.3102/0034654309332490

King, J., Brundiers, K., & Fischer, D. (2024). Student agency in a sustainability-oriented assessment process: exploring expansive learning in student-led rubric co-design. *Assessment & Evaluation in Higher Education*, 1-13. https://doi.org/10.1080/02602938.2024.2333031

Kobrin, J. (2022). Evidence-centered design. In B.B. Frey (Ed.), *The SAGE Encyclopedia of Research Design* (2nd ed., Vol. 4) (pp. 514-516). SAGE Publications, Inc. https://doi.org/10.4135/9781071812082

Koran, J., & Kopriva, R.J. (2017). Framing appropriate accommodations in terms of individual need: Examining the fit of four approaches to selecting accommodations of English language learners. *Applied Measurement in Education, 30*(2), 71-81. https://doi.org/10.1080/08957347.2016.1243539

Kusters, A., Spotti, M., Swanwick, R., & Tapio, E. (2017). Beyond languages, beyond modalities: Transforming the study of semiotic repertoires. *International Journal of Multilingualism, 14*(3), 219–232. https://doi.org/10.1080/14790718.2017.1321651

Li, H., & Suen, H.K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education, 25*(4), 327-346. https://doi.org/10.1080/14790718.2017.1321651

Li, W. (2011). Moment analysis and translanguaging space: Discursive construction of identities by multilingual Chinese youth in Britain. *Journal of Pragmatics, 43*, 1222–1235. https://doi.org/10.1080/14790718.2017.1321651

Liu, L. (2023). Accommodations for English language learners in large-scale assessments: Testing accommodations for English language learners. *The Educational Review, 7*(6), 688–692. http://dx.doi.org/10.26855/er.2023.06.006

López, A.A. (2023). Examining how Spanish-speaking English language learners use their linguistic resources and language modes in a dual language mathematics assessment task.

*Journal of Latinos and Education, 22*(1), 198-210. https://doi.org/10.1080/15348431.2020.1731693

López, A.A., Guzman-Orth, D.A., & Turkan, S. (2015). How might a translanguaging approach in assessment make tests more valid and fair for emergent bilinguals? In G. Valdés, K. Menken & M. Castro, M. (Eds.), *Common core, bilingual and English language learners: A resource for educators* (pp. 266–267). Caslon.

López, A.A., Guzman-Orth, D., & Turkan, S. (2019). Exploring the use of translanguaging to measure the mathematics knowledge of emergent bilingual students. *Translation and Translanguaging in Multilingual Contexts, 5*(2), 143-164. https://doi.org/10.1075/ttmc.00029.lop

López, A.A., Turkan, S., & Guzman-Orth, D. (2017). Conceptualizing the use of translanguaging in initial content assessments for newly arrived emergent bilingual students. *ETS Research Report, 17*(7). https://doi.org/10.1002/ets2.12140

Lyon, E.G. (2023). Reframing formative assessment for emergent bilinguals: Linguistically responsive assessing in science classrooms. *Science Education, 107*(1), 203–233. https://doi.org/10.1002/sce.21760

Martin-Beltrán, M. (2014). "What do you want to say?" How adolescents use translanguaging to expand learning opportunities. *International Multilingual Research Journal, 8*(3), 208–230. https://doi.org/10.1080/19313152.2014.914372

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment, 14*(3–4), 160–79. https://doi.org/10.1080/10627190903422906

Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). A brief introduction to evidence-centered design. *ETS Research Report, 13*(16). https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review, 6*(3), 281–307. https://doi.org/10.1515/applirev-2015-0014

Otheguy, R., García, O., & Reid, W. (2019). A translanguaging view of the linguistic system of bilinguals. *Applied Linguistics Review, 10*(4), 625-651. https://doi.org/10.1515/applirev-2018-0020

Paradis, J., Genesee, F., & Crago, M.B. (2010). *Dual language development and disorders: A handbook on bilingualism and second language learning* (2nd Ed.). Paul H. Brookes.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practices, 30*(3), 10–28. https://doi.org/10.1111/j.1745-3992.2011.00207.x

Rios, J.A., Ihlenfeldt, & Chavez, C. (2020). Are accommodations for English learners on state accountability assessments evidence-based? A multistudy systematic review and meta-analysis. *Educational Measurement: Issues and Practices, 39*(4), 65-75. https://doi.org/10.1111/emip.12337

Rivera, C., & Stansfield, C.W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment, 9*(3-4), 79-105. https://doi.org/10.1080/10627197.2004.9652960

Roohr, K.C. & Sireci, S.G. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment, 22*(1), 35-53. https://doi.org/10.1080/10627197.2016.1271704

Roschmann, S., Witmer, S.E., & Volker, M.A. (2021). Examining provision and sufficiency of testing accommodations for English learners. *International Journal of Testing, 21*(1), 32–55. https://doi.org/10.1080/15305058.2021.1884872

Sanchez, S.V., Rodriguez, B.J., Soto-Huerta, M.E., Castro Villareal, F., Guerra, N.S., & Bustos Flores, B. (2013). A case for multidimensional bilingual assessment. *Language Assessment Quarterly, 10*(3), 160–177. https://doi.org/10.1080/15434303.2013.769544

Sayer, P. (2013). Translanguaging, TexMex, and bilingual pedagogy: Emergent bilinguals learning through the vernacular. *TESOL Quarterly*, *47*(1), 63-88. https://doi.org/10.1002/tesq.53

Shohamy, E. (2011). Assessing multilingual competencies: Adopting construct valid assessment policies. *The Modern Language Journal*, *95*(3), 418-429. https://doi.org/10.1111/j.1540-4781.2011.01210.x

Smarter Balanced Assessment Consortium (2012). *Translation accommodations framework for testing English language learners in Mathematics*. Retrieved from https://portal.smarterbalanced.org/library/en/translation-accommodations-framework-for-testing-english-language-learners-in-mathematics.pdf

Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher*, *37*(4), 189–199. https://doi.org/10.3102/0013189X08319569

Turkan, S., López, A.A., Lawless, R.R., & Tolentino, F. (2019). Using pictorial glossaries as an accommodation for English learners: An exploratory study. *Educational Assessment, 24*(3), 235–265. https://doi.org/10.1080/10627197.2019.1615371

Velasco, P., & García, O. (2014). Translanguaging and the writing of bilingual learners. *Bilingual Research Journal*, *37*(1), 6–23. https://doi.org/10.1080/15235882.2014.893270

Walker, M.E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). *Culturally responsive assessment: Provisional principles*. *ETS Research Report, 23*(11). https://doi.org/10.1002/ets2.12374

Wise, S.L., & DeMars, C.E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wolf, M.K., Kim, J., & Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Applied Measurement in Education, 25*(4), 347-374. https://doi.org/10.1080/08957347.2012.714693

Wolf, M.K., Yoo, H., Guzman-Orth, D., & Abedi, J. (2021). Investigating the effects of test accommodations with process data for English learners in a mathematics assessment. *Educational Assessment, 27*(1), 27–45. https://doi.org/10.1080/10627197.2021.1982693

Yang, X. (2024). Linguistically responsive formative assessment for emergent bilinguals: exploration of an elementary teacher's practice in a math classroom. *International Multilingual Research Journal*, 1–24. https://doi.org/10.1080/19313152.2024.2339757