e-ISSN: 2148-7456

a peer-reviewed
online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 10 Issue: 4 December 2023

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

e-ISSN: 2148-7456

Volume 10 Issue 4 2023

Editor Dr. Omer KUTLU

Address Ankara University, Faculty of Educational Sciences, Cebeci Yerleşkesi, Çankaya, Ankara, Türkiye

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

Publisher Dr. Izzet KARA

Address Pamukkale University, Education Faculty, Kinikli Campus, 20070 Denizli, Türkiye

Phone +90 258 296 1036

Fax +90 258 296 1200

E-mail ikara@pau.edu.tr
ijate.editor@gmail.com

Journal Contact Dr. Eren Can AYBEK

Address Department of Educational Sciences, Pamukkale University, Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye

Phone +90 258 296 31050

Fax +90 258 296 1200

E-mail erencanaybek@gmail.com

Address Dr. Anil KANDEMIR
Department of Educational Sciences, Agri Ibrahim Cecen University, Faculty of Education, Agri, Türkiye
akandemir@agri.edu.tr

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

Online ISSN 2148-7456

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

Cover Design Merve SENTURK

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

International Journal of Assessment Tools in Education

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

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

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

Indexes and Platforms:

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

CONTENTS

Research Articles

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 638–671

https://doi.org/10.21449/ijate.1243812

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Adaptation of an entrepreneurship education self-assessment scale at the tertiary level into Turkish

Ahmet Celik[1,*], **Ebru Solmaz**[1]

[1]Gazi University, Distance Education Application and Research Center, Ankara, Türkiye

ARTICLE HISTORY

Received: Jan. 28, 2023
Revised: Aug. 25, 2023
Accepted: Sep. 19, 2023

Keywords:
Higher education,
Entrepreneurship education,
Entrepreneurial university,
Self-assessment,
Scale adaptation.

Abstract: The university environment provides a good context for entrepreneurship education. With the vigorous development of entrepreneurship education, educators and scholars have shown increasing interest in the significant role entrepreneurship education plays in higher education. As a result, the effectiveness of entrepreneurship education has quickly become a popular topic. However, it is often not easy to evaluate entrepreneurship education programs, which are designed for medium- and long-term outcomes. It is essential to develop alternative assessment tools that do not traditionally assess only knowledge. The study aims to adapt the multidimensional measurement tool for assessing university students' entrepreneurial skills, knowledge, attitudes, and mindsets. While the scale was translated into Turkish, face and content validity were proved. The data was gathered from 572 university students. Confirmatory factor analyses were employed to assess the construct validity of the measure. The Turkish Entrepreneurship Education Self-Efficiency Scale was obtained with three main dimensions and 38 items. Its Cronbach's alpha, Spearman-Brown correlation, and composite reliability coefficients are 0.95, 0.86, and 0.98, respectively. Furthermore, the study found that the entrepreneurship education scores of the participants were significantly related to their gender, field of education, volunteering, work experience, experience of starting or running their own business, and entrepreneurship education. The effect size of these variables differs, and the experience with self-employment has the greatest influence on entrepreneurship education.

1. INTRODUCTION

Given the current economic challenges and the recession in the global economy with the COVID-19 pandemic facing countries worldwide, creating more widespread and effective entrepreneurial activities has become an important goal. The economy, labour markets, societies, and social structures are increasingly undergoing continuous change as an effect of globalization. As lifelong learning has become necessary for all citizens, we need to develop throughout our lives not only for personal development but also our ability to shape the society we live in and our skills to succeed in an ever-changing world (European Council, 2018). This environment of rapid change created by Industry 4.0 and globalization means that anyone should have specific knowledge and skills related to their work and entrepreneurial competence

*CONTACT: Ahmet CELİK ✉ ahmetcelik@gazi.edu.tr Gazi University, Distance Education Application and Research Center, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

that will enable them to adapt to uncertainty. To cope with this constant change, it has become an important goal for all countries to make individuals competent to function as entrepreneurs.

To achieve this goal, it is essential that the education systems focus on developing entrepreneurial skills and capabilities from primary to higher education. It is now well-recognized that education and training opportunities play a key role in developing future entrepreneurs (Henry et al., 2005). In this sense, entrepreneurship education promotes the entrepreneurial mindset by providing students with skill sets, knowledge, and behavioral patterns that enable them to become entrepreneurs in their own lives (Moberg et al., 2009). However, because entrepreneurial skills are seen as synonymous with starting a business, and entrepreneurship education policies vary from country to country, there are many ways of evaluating entrepreneurship, especially in higher education institutions. The highly complex relationship between the concepts of entrepreneurship and education leads to problems in evaluating the impact of entrepreneurship education at the tertiary level, especially in the Republic of Turkey, using traditional methods. This article examines the current and valid definition of entrepreneurship education based on literature. Then it aims to adapt a scale developed in multiple languages by EU countries, which can assess the impact of entrepreneurship education at the tertiary level in a multidimensional way into Turkish.

1.1. A Quick Overview of Entrepreneurship

The term entrepreneurship comes from the French word entrepreneur, rooted in the word enterprise, meaning someone who undertakes a business. The word entrepreneur is also often used to mean "someone who takes on the risks and management of a business". This concept, which originated earlier, became popularized with the First Industrial Revolution in the 19th century and has been associated with business establishment and business activities since then. However, most business and finance-oriented definitions of entrepreneurship shaped by this trend are about starting a business or assuming the risks associated with running a business. Therefore, these definitions emphasize entrepreneurship's ability to transform any industry and scale solutions faster than other economic approaches (Shamsrizi et al., 2021), with entrepreneurship being an important economic growth driver providing national advantage (Porter, 1990).

Entrepreneurship defines a broader process beyond starting a company in the Shane and Venkataraman (2000) model of entrepreneurship. In defining entrepreneurship, Venkataraman (1997) emphasizes the processes of discovering, evaluating, and exploiting opportunities to create value. Entrepreneurship is therefore, strongly associated with the ability to recognize and exploit opportunities in the environment (Gibb, 2005). Different studies in the field of entrepreneurship describe in detail how entrepreneurial skills are applied in various processes of entrepreneurship, such as identifying opportunities, resolving conflicts, and dealing with uncertainty (Malywanga et al., 2020). These skills are deployed at specific stages of entrepreneurship, which should be understood and studied as a cognitive and evolving process rather than a one-time decision (Arkko-Saukkonen, 2017).

In their study on the entrepreneurial process, van der Veen and Wakkee (2016) argued that focusing only on the characteristics of the entrepreneur does not reflect the realities of the entrepreneurial process. Entrepreneurship is not about entrepreneurs' psychology and character traits, but about their actions, behaviours, and related concepts (Drucker, 1985a, 1985b). According to O'Hara (2011)'s study on the relationship between entrepreneurial skills and entrepreneurial processes, 4 key behaviours are prominent in entrepreneurship (as cited in Cooney, 2012):

- The ability to identify and exploit a business opportunity.
- The human creative endeavour of developing a business or building something of value.

- Willingness to take risks.
- The ability to organize the resources necessary to respond to the opportunity.

These behaviours embody that entrepreneurship is more important than taking high risks; it is about being able to manage the business process. The effort to explain entrepreneurship by emphasizing its inherent behaviours, i.e., the process, rather than a genetic inheritance or a result of personality traits, is vital in showing that entrepreneurship can be learned through education like any other discipline (Alum, 1986; Chimucheka, 2014; Drucker, 1982, p. 143). As a result of this effort, there has recently been a growing awareness worldwide, especially in the European Union, that people's entrepreneurial competencies can be developed through learning. The European Commission first identified 'a sense of initiative and entrepreneurship' as one of the eight core competences required for all members of a knowledge-based society (European Commission, 2007). After a decade, the European Commission (2019) has developed a lifelong learning competences framework that aims to create a common understanding covering a wide range of learning environments from primary school to university.

1.2. Entrepreneurship Education

Education has played a crucial part in developing entrepreneurial competencies in individuals and shaping their inclination towards exhibiting entrepreneurial behaviors. While education alone may not be adequate to enhance the tendency towards entrepreneurial endeavors (Balaban & Özdemir, 2008), it does hold significant significance in fostering the development and continuity of an entrepreneurial culture within society (European Commission, 2012; Genç, 2019). Entrepreneurship education has been found to have positive social effects, as evidenced by research conducted by the European Commission (2012) and Fayolle et al. (2006). This form of education has been shown to foster individuals' aspirations to become entrepreneurs promoting their ambition and engagement in extracurricular activities that contribute to their personal growth. Moreover, entrepreneurship education has been found to enhance individuals' awareness of innovation, bolster their communication skills, and heighten their motivation towards entrepreneurship, as indicated by studies conducted by Cevher (2016), Nasr and Boujelbene (2014), and Uygun et al. (2018). At the economic level, entrepreneurship education has been shown to increase the number of business start-ups, create new jobs and raise taxable income (Elert et al., 2015; Martin et al., 2013).

The main goals of entrepreneurship education encompass equipping young individuals with specific knowledge, skills, and attitudes to cultivate an entrepreneurial mindset, thereby fostering entrepreneurial behavior in their personal lives (Moberg et al., 2009). Additionally, this form of education aims to enhance their creativity and self-assurance in their endeavors and contributions to society and the economy (European Commission, 2012). A recent study highlights the importance of incorporating entrepreneurship education into the curriculum, starting from early schooling, and continuing through higher education. Early entrepreneurship education is seen as a means to cultivate persons with entrepreneurial skills who are equipped to navigate the intricate difficulties of the 21st century (European Council, 2013). Universities, specifically, have a significant responsibility to fulfill in augmenting these endeavors for the youth.

Today, universities are institutions that determine the types and fields of their activities according to the needs and expectations of global society and the resources allocated to them (Yelkikalan et al., 2010). Increasing economic integration, global competition, and the development of new information and communication technologies have led universities to assume a role that will directly contribute to economic and social development (Sakınç & Bursalıoğlu, 2012). With the addition of contribution to economic and social development to education and research activities, universities must have an entrepreneurial organizational structure that supports change-oriented activities to provide individuals with opportunities to adapt to an ever-changing society. In an entrepreneurial university, norms, values, and expectations support entrepreneurship, and people engage in entrepreneurial activities (Sherkat & Chenari, 2020). However, another factor as crucial as the entrepreneurial structure of universities is the role, they play in developing the entrepreneurial competencies of individuals through the dissemination of entrepreneurship education (Yelkikalan et al., 2010).

From the view of structuration theory, competencies are developed over time (Morris et al., 2013), and university education can play an important role in this process by providing real-world opportunities. According to Schulte (2007), developing students' entrepreneurial spirit in all areas is among the main tasks of an entrepreneurial university. The ability of a university to educate its graduates not only as job seekers but also as entrepreneurs who can create jobs is one of the most critical drivers of entrepreneurship. Empirical research has shown that entrepreneurship education at universities positively promotes entrepreneurial attitudes and develops young graduates' human capital (Aboobaker & Renjini, 2020; Johannisson, 2006; Roman & Maxim, 2017; Varela & Jimenez, 2001). Therefore, to develop the entrepreneurial university potential, having a holistic approach that will create synergy between the university's entrepreneurship activities and entrepreneurship education is valuable. According to Gibb (2012), entrepreneurship education is one of the 5 key areas that provide this potential and synergy of universities.

According to a study among graduates of higher education institutions in Europe, the entrepreneurship education young people receive at university has a positive impact on their entrepreneurial mindset, their entrepreneurship intentions, their degree of taking the initiative, their employability, and ultimately their role in society and the economy (European Commission, 2012). Therefore, entrepreneurship education is expected to improve not only the role of the individual in the economy but also his/her social and personal life in society. Entrepreneurship education influences students' future intentions, and becoming an entrepreneur is among the common career plans of university students after graduation (Rasmussen & Sørheim, 2006). Many universities want to expand entrepreneurship courses to contribute positively to this tendency of students (Galloway & Brown, 2002; Henderson & Robertson, 1999).

In Turkey, compulsory or elective entrepreneurship courses taught directly in public and foundation universities and entrepreneurship trainings led by the Small and Medium Enterprises Development Organization of Turkey are examples of this effort (Genç, 2019). However, there are also general trainings in which entrepreneurship culture is embedded in higher education through courses, research, consultancies, and all other activities at the institutional level in other countries (McMullan & Gillin, 1998; Rasmussen & Sørheim, 2006). In Turkey, only a small number of universities have such trainings. As in many European universities, entrepreneurship trainings offered through specific courses under the economics and business administration departments are more common in Turkey. In addition, European Union (EU) countries are developing national strategies and action plans for entrepreneurship education, structuring curricula, and developing practices to support teachers in line with the decisions of the "European Reference Framework for Key Competences in Lifelong Learning" (European Commission, 2007; European Council, 2006), key competences framework for lifelong learning (European Commission, 2016; European Council, 2018), and the "2020 Entrepreneurship Action Plan" (European Council, 2013). Unlike EU countries, Turkey has several outworn strategies related to entrepreneurship education, the most relevant of which is the 'Ministry of National Education Strategic Plan 2010-2014'. These strategies do not include any monitoring and evaluation plan (European Commission/EACEA/Eurydice, 2016).

Just as important as having entrepreneurship education strategies in higher education institutions is monitoring, measuring, and evaluating practices to determine whether these strategies achieve their objectives. Even EU member states, which have made significant progress in entrepreneurship education to date, are still considering how to measure the impact of the national entrepreneurship education strategies they have implemented at the policy level (European Commission, 2012). The European Commission report 2001 clearly stated that one of the biggest problems with entrepreneurship education is the inadequacies in evaluating entrepreneurship education (Andrijevskaja & Mets, 2008). Therefore, assessment and evaluation practices as an element of entrepreneurship education program design are seen as an obstacle that needs to be overcome to embed entrepreneurship education in tertiary level curricula effectively (European Commission/EACEA/Eurydice, 2016; Fayolle et al., 2006).

1.2.1. *Problems in assessing entrepreneurship education at tertiary level*

As the interest of educators and scholars in the role of entrepreneurship education in higher education has increased, a wide variety of definitions, objectives, content, and pedagogical methods have arisen (Fayolle, 2008; Liu et al., 2021). Given this lack of standardization, assessment becomes fundamental in improving the effectiveness and efficiency of entrepreneurship education (Béchard & Grégoire, 2005), and accurately assessing entrepreneurship education has quickly become a popular topic. Although assessing the impact of entrepreneurship education may seem complicated because it must include many types, purposes, and methods of assessment, it is beneficial because it provides an opportunity for program improvement (Fayolle & Gailly, 2015; Galvão et al., 2019; McMullan & Gillin, 1998). However, the academic challenges in evaluating entrepreneurship education programs (Fayolle et al., 2006); the neglect of real outcomes that entrepreneurs need when evaluating the effectiveness of entrepreneurship education in previous research (Scott et al., 2016); and the need to evaluate programs (Garavan & O'Cinneide, 1994; Honig, 2004; Pittaway & Edwards, 2012) call for more research on this topic.

Liu et al. (2021) explained this deficiency in the literature by addressing it in two dimensions. The first dimension is the lack of general validity arising from using many different indicators separately in evaluating the effectiveness of entrepreneurship education. Although changes in the selected indicators seem to reflect the impact of entrepreneurship education, the validity of the evaluation with a single indicator is relatively low. Secondly, there is a lack of a unified measurement model under a standard framework. A framework of multiple indicators is needed to reduce the limitations of unidimensional instruments and scales in comparative cross-regional, cross-cultural, and cross-institutional entrepreneurship education studies. However, the existing literature also lacks studies that explore the logical relationships between multiple indicators. This is because it is often not easy to evaluate entrepreneurship education programs, which by their very nature are designed for medium and long-term outcomes (McMullan & Gillin, 1998).

Pittaway et al. (2009) reviewed the literature and observed that although most entrepreneurship education research focuses on program design and implementation, there is a significant gap in evaluation practices. Fayolle et al. (2006) assessed the social, cultural, and economic impacts of entrepreneurship education from a new perspective using an evaluation approach based on the theory of planned behaviour to overcome the uncertainties in the selection of criteria identified in the literature. Jones and Penaluna (2013) found that being overly prescriptive in assessment strategies limited students' achievement of the targeted entrepreneurial competencies; therefore, they argued that assessment practices should be more flexible, more accepting of ambiguity, and formative in nature. Sherkat and Chenari (2020) used a model previously proposed by Fayolle and Gailly (2008) to assess different components of entrepreneurship education. Liu et al. (2021), who wanted to eliminate the problems of single indicator-based measurements, stated that they were able to comprehensively measure the impact of entrepreneurship education for university students by using a model consisting of the dimensions of entrepreneurial competencies, perception of entrepreneurial barriers, and entrepreneurial intentions.

The fact that entrepreneurship education programs in higher education institutions are offered in different types, integrated into the core curriculum or stand-alone, with more comprehensive levels, including courses, multiple courses, or institution-wide experiential learning, highlights the complexity associated with the assessment of entrepreneurship education. This complexity is further compounded by the fact that assessment is also driven by the need to support students' progress (formative assessment) and determine student performance (summative assessment) to meet the requirements for certified accreditation. In this context, it is necessary to develop feasible assessment practices within educational processes to monitor the impact of entrepreneurship education in higher education (Pittaway & Edwards, 2012). Identifying validated instruments that could measure the scope of entrepreneurship education outcomes is a major challenge (Duval-Couetil et al., 2010).

1.2.2. *Assessing entrepreneurship education with a multi-dimensional scale*

Developing instruments to measure different psychological constructs, such as entrepreneurial self-efficacy (ESE), entrepreneurial orientation, and entrepreneurial intention is a hot topic in business, management, and education. Some studies assessed students' perceptions of business skills and knowledge, self-efficacy, attitude towards entrepreneurship, entrepreneurial intent (Huang-Saad et al., 2016), and students' level of interest in entrepreneurial education (Shinnar et al., 2009) with different scales. In another study, competencies defined specific to the entrepreneurship discipline to develop scales to measure the effectiveness of entrepreneurship education (Morris et al., 2013). Due to its relevance with business, the competency-based approach has thus far established a standard paradigm for this kind of research (Mitchelmore & Rowley, 2010). Most studies, however, are theoretical works, and those that do offer empirical data instead identify entrepreneurship abilities without providing a strong theoretical foundation (Silveyra et al., 2021). Thus, these studies are not entirely applicable to assessing entrepreneurship education offered in different disciplines at the higher education level.

There are a few studies to develop various assessment tools published in the literature to assess university students' entrepreneurship education through multi-dimensional outputs (Bamiatzi et al., 2015; Duval-Couetil et al., 2010; Silveyra et al., 2021). In several studies, the effectiveness of entrepreneurship education has been assessed by focusing solely on a specific entrepreneurial behaviour. Saeed et al. (2014), for instance, used multiple scales to create a questionnaire and introduced a multi-level perspective of the factors that influence entrepreneurial intention. It has been observed that scales from different disciplines or questionnaire-based surveys are utilized in other research with a similar goal of evaluating entrepreneurship education, and in some cases, these scales have even undergone revisions following the research questions (Ahmed et al., 2017; Hasan et al., 2017; Mitchelmore & Rowley, 2013; Vanevenhoven & Liguori, 2013). However, most of these studies seem to be designed for research and scholarly study at a particular point in time, rather than for ongoing course or program evaluation (Duval-Couetil, 2013).

In a recent review of key studies conducted in the field over the last 15 years, Rideout (2012) noted that while progress has been made, scientific knowledge on the evaluation of entrepreneurship programs remains at an early stage of development and has a long way to go before the field can confidently answer the questions of whether and how entrepreneurship education works. Given that few of the studies conducted and evaluation tools developed have been validated through replication or used in multiple contexts or populations, it is clear that to overcome the difficulties of assessment in entrepreneurship education, universities need practical and accessible measurement tools that can assess the impact of education in terms of entrepreneurship.

Given that entrepreneurship education is meant to equip people with the knowledge, skills, and attitudes to act entrepreneurially throughout their lives to determine the extent to which this goal is being achieved throughout the educational life of students in higher education institutions, it is essential to develop measurement tools that reflect the fact that entrepreneurship is a crucial competence for life and has importance far beyond simply aiming to start a business. On the other hand, there is also a need to organize and validate conventional indicators to ensure consistency and comparability of results after assessing entrepreneurial knowledge, skills, and attitudes together. In addition, since self-assessment and peer assessment are not used as often as expected in entrepreneurship education evaluation practices, assessment studies are regrettably limited to traditional methods (Pittaway & Edwards, 2012).

Entrepreneurship education ranks high on European policy agendas, but little research is available to assess its impact (von Graevenitz et al., 2010). Entrepreneurship education is typically offered as an elective course at Turkish institutions. Even though these courses are in great demand, students frequently enroll in them with the expectation of succeeding (Marangoz & Taçyu Dolu, 2022). Due to this circumstance, research in Turkey has a high tendency to evaluate the immediate effects of specific educational activities using a wide range of scales (Pazarcık, 2016). As a result, there is very little chance of comparing different study outcomes.

It will be meaningful to use valid and reliable scales that (1) define entrepreneurship as a multidimensional competence consisting of knowledge, skills, attitudes, and behaviours, (2) view entrepreneurship education as processes designed to develop this competence, (3) are applicable and valuable in all disciplines and lines of education, (4) are suitable for self-assessment in the form of pre and post-test to reveal the effect of education, (5) can offer comprehensive suggestions for university management and policy makers to develop entrepreneurship from a multidimensional perspective and to determine the effect of entrepreneurship education, which is given as a whole course or education program in universities in Turkey. One of the studies to validate and develop a measurement tool at higher education to help close this gap was carried out by Moberg et al. (2009) in an EU project called "Assessment Tools and Indicators for Entrepreneurship Education (ASTEE)."

The project started because of the need to measure the influence of entrepreneurship education at all levels of education among pupils and students (primary, secondary, tertiary) to improve and promote the dissemination of entrepreneurship education by providing educational institutions in Europe with access to these tools going forward (The ENTREDU, n.d.). Therefore, a project consortium of EU Member States (Ireland, France, Portugal, Germany, Croatia, Belgium, and Denmark) was established to develop a common indicator framework and measurement tool that could be used across EU countries. It was stated that these countries were selected because they represent the EU very well with different levels of maturity of entrepreneurship education practices at all levels of education systems (Moberg et al., 2009). In the project coordinated by the Danish Foundation for Entrepreneurship - Young Enterprise, in addition to the partners, Sweden, the United Kingdom, Austria, Italy, Romania, and Spain were also involved in the development and testing processes to increase the applicability of the measurement tools across Europe (Moberg et al., 2009; OECD/European Union, 2018). Thus, differences in education systems in the EU, cultural differences, and views on entrepreneurship education were aimed to be reflected in the scale as much as possible.

The tertiary level scale developed in the ASTEE project can be used as a self-assessment tool to determine the level of entrepreneurial competencies of university students participating in entrepreneurship courses, entrepreneurship education embedded in a specific subject, course, or discipline, or general education by age group. It can also be used by teachers, educators, policymakers, and researchers before and after training to measure how students' entrepreneurial competences are improved through educational content and methods (ASTEE User Guide, n.d.).

In recent years, there has been a growing interest among researchers in Turkey regarding the concept of entrepreneurship. Consequently, there has been a noticeable increase in studies focusing on entrepreneurship education within universities (Balaban & Özdemir, 2008; Bulut & Aslan, 2014; Bozkurt & Alparslan, 2013; Çolakoğlu & Çolakoğlu, 2016; Özdemir, 2016; Uygun & Güner, 2016). When analyzing the research conducted on entrepreneurship education, it is evident that surveys are commonly employed as the primary method for data collecting (Balaban & Özdemir, 2008; Çolakoğlu & Çolakoğlu, 2016; Uygun & Güner, 2016). Nevertheless, despite the absence of a competency-based standardized measurement tool assessing the impact of entrepreneurship education in the domestic literature, Yılmaz and Sünbül (2009) have devised a scale to evaluate the entrepreneurship levels of university students. The one-dimensional scale, including 36 items, was subjected to a study of its validity and reliability among students enrolled in a university faculty. In addition, Sart (2020) established a scale to measure individual entrepreneurial tendencies at the university level, including five sub-dimensions and a total of 30 questions. In a similar vein, Ercan and Yıldıran (2021) undertook adapting the individual entrepreneurial tendency scale for university students into the Turkish language. The scale is composed of three dimensions and encompasses ten items. In this situation, it could be argued that using a multidimensional measurement tool from Turkey that measures the effect of entrepreneurship education while considering its impact on making people more entrepreneurial would be a great way to find and implement effective educational procedures.

In light of the above information, considering the developments and needs in assessment and evaluation studies on entrepreneurship education, it is thought that a valid and reliable measurement tool adapted to the Turkish language and the cultural structure of the country will contribute to an essential need in the field of higher education. This study aimed to adapt the tertiary level scale developed in the ASTEE project for EU member states into the Turkish language to be used in Turkey, one of the EU candidate countries. Thus, it will contribute to create effective, comprehensive, and generalizable measurement processes in studies aiming to measure and evaluate the impact of entrepreneurship education in higher education in Turkey and to increase the comparability of the results obtained with international literature, especially with EU countries. In this context, validity and reliability analyses of the scale translated into the Turkish language will be carried out, and a valid and reliable measurement tool that can evaluate the impact of entrepreneurship education that students receive in the context of university education will be presented to the use of policy makers, administrators, lecturers, and field experts. Moreover, many individual variables can affect entrepreneurship competency at different levels in different cultures. In this research, it is also aimed to reveal the effects of these variables on entrepreneurship education in Turkish culture.

2. METHOD

2.1. Participants

The study was conducted with 582 undergraduate students studying in public and foundation universities in Turkey in the 2018-2019 and 2020-2021 academic years. The data of 10 individuals with excessive and missing values were excluded, and the validity and reliability analyses of the study were completed with the data obtained from 572 participants. The demographic information of the participants is given in Table 1.

Table 1. *Demographic information of participants.*

	Group	n	%
Gender	Female	364	63.6%
	Male	208	36,4%
University	State	562	98.3%
	Foundation	10	1.7%
Duration of study in higher education	1 year	65	11.4%
	2 years	39	6.8%
	3 years	195	34.1%
	4 years	186	32.5%
	5 years	67	11.7%
	6 years	9	1.6%
	Over 6 years	11	1.9%
Entrepreneurship Education Experience	Yes	202	35.3%
	No	370	64.7%
Experience of starting/running own business	Yes	191	33.4%
	No	381	66.6%

The study group consisted of 63.6% (n=364) female and 36.4% (n=208) male undergraduate students. Of the participants, 98.3% (n=562) were studying at state universities and 1.7% (n=10) at foundation universities in Turkey. While data were collected from a total of 25 universities, 21 of these universities were public universities, and 4 were private universities. According to the demographic data on higher education experience, 34.1% (n=195) of the students had 3 years of experience; 32.5% (n=186) had 4 years of experience; 11.7% (n=67) had 5 years of experience; 11.4% (n=65) had 1 year of experience; 6.8% (n=39) had 2 years of experience; 1.9% (n=11) had more than 6 years of experience; and 1.6% (n=9) had 6 years of experience. Since the scale focuses on skills and competencies for entrepreneurship education, participants asked about experience in entrepreneurship education and starting/running their own business. According to this data, 64.7% (n=370) of the participants were not taking or had not taken an entrepreneurship course out of school or at school, while 35.3% (n=202) were currently taking a course or had taken one in the past. At the same time, 33.4% (n=191) of the participants have experience starting/running their own business, while 66.6% (n=381) do not.

2.2. Measurement Tool

Entrepreneurship education assessment tool for the tertiary level developed in English by the ASTEE project partners was adapted into Turkish in this study. The Turkish version is called "Entrepreneurship Education Self-Assessment Scale (EESS)." The original scale, which aims to develop measurement tools to assess the entrepreneurship skills, knowledge, attitudes, and mindsets of higher education students, was developed between December 2012 and June 2014 for students over the age of 20. The scale consists of 57 items, including 18 open and closed-ended questions on demographic information and 39 Likert-type questions on entrepreneurial competencies. The items related to entrepreneurial competence, which are collected in 3 main dimensions and 11 sub-dimensions under the titles of *"mindset"*, *"skills - entrepreneurial self-efficacy (ESE)"*, and *"career ambitions"* are scored between 1-7.

The "Mindset" dimension of the scale includes 11 items in 3 sub-dimensions: the "entrepreneurial mindset" with 3 items, the "core self-evaluation" sub-dimension with 5 items, and the "entrepreneurial attitudes" sub-dimension with 3 items. The "Entrepreneurial skills" dimension consists of 22 items in a total of 6 sub-dimensions: the "creativity" sub-dimension with 4 items, the "planning" sub-dimension with 4 items, the "financial literacy" sub-dimension with 3 items, the "marshalling resources" sub-dimension with 4 items, the "managing uncertainty" sub-dimension with 4 items, and the "entrepreneurial knowledge" sub-dimension with 3 items. The "Career Ambitions" dimension consists of 6 items in 2 dimensions: the "innovative employee" sub-dimension with 3 items and the "entrepreneurial intentions" sub-dimension with 3 items (Moberg et al., 2009). The distribution of all dimensions of the scale and the number of items, and some of the items are presented in Table 2.

Table 2. *Original ASTEE tertiary level measurement tool factors and items.*

Factors	Sub-factors	Number of Items	Examples of Items	Reliability
Mindset	Entrepreneurial mindset	3	I am often the first one to suggest a solution to a problem	.73
	Core self-evaluation	5	When I try, I generally succeed	.88
	Entrepreneurial attitudes	3	*In general, starting a business is…* Negative / positive	*.87*
ESE (Skills)	Creativity	4	*I am able to…* Identify opportunities for new ways to conduct activities	*.84*
	Planning	4	*I am able to….* Network (i.e., make contacts with and exchange information with others)	*.86*
	Financial literacy	3	*I am able to…* Control costs for projects	*.80*
	Marshalling of resources	4	*I am able to…* Put together the right group/team in order to solve a problem	*.85*
	Managing ambiguity	4	*I am able to…* Manage uncertainty in projects and processes	*.80*
	Entrepreneurial Knowledge	3	That some business ideas work, and others don't	.85
Career Ambitions	Innovative employee	3	*I would like to have a job that allows me to...* Solve problems in new ways	*.85*
	Entrepreneurial intentions	3	I have business ideas I am going to implement	.91
Total		39		

As a result of confirmatory factor analysis in the original (English) scale, chi-square/*df* was calculated as 3.5; RMSEA .051, CFI .994, TLI .933. The coefficients of reliability were .73 for "Entrepreneurial Mindset", .88 for "Core-self Evaluation", and .87 for "Entrepreneurial Attitudes" in the "Mindset" factor; .84 for "Creativity", .86 for "Planning, .80 for "Financial Literacy"; .85 for "Marshalling of Resources", .80 for "Managing ambiguity" and .85 for "Entrepreneurial knowledge" in "ESE Skills" factor; .91 for "Entrepreneurial intentions" and .85 for "Innovative employee" in "Career Ambitions" factor.

2.3. Translation Process

In this study, the evidence for different types of validity were collected during the adaptation process of the scale into Turkish. Before starting, the researchers obtained permission from the project managers to use the scale. Then a total of fifty-seven items, including demographic questions, were translated into Turkish by the researchers. According to Seçer (2018), rather than translating the items into Turkish one-to-one, the items should be examined by Turkish and foreign language experts to adapt the scale culturally and making the necessary adjustments contributes to validity and reliability. In this direction, the translated scale was sent to 5 entrepreneurship education experts and 3 English language experts, and they were asked to evaluate the translations together with the original scale. Necessary corrections were made in line with the suggestions received from the experts, and the experts were asked to check the scale for the second time.

Review of the literature and expert comments were used to gather evidence of the scale's content validity. According to the expert reviews, the measurement tool serves its goal because it deals with entrepreneurial competencies and entrepreneurial behaviours that should be focused on at the university level. Furthermore, a review of the entrepreneurship education literature reveals that debates centered on individual characteristics have given way to research centered on the learning elements of entrepreneurial behaviour. Interestingly, recent research has given particular attention to how people develop their competence for creativity, opportunity recognition, resource management, and initiative. Kyrö (2006) drew attention to this link between individual and career and then underlined the need to consider both the development of an individual's potential and entrepreneurial behaviours while defining the elements of entrepreneurship education. Heinonen and Poikkijoki (2006) asserted that entrepreneurship education entails the development of an individual's knowledge, abilities, and attitudes. As a result of the competence attained from these elements, the individual can use entrepreneurship in her career and personal life. These elements are characterized by three interrelated dimensions in the EntreComp conceptual framework: ideas, resources, and activities. (Bacigalupo et al., 2016). In another study, learning elements of entrepreneurship were classified into five dimensions with a taxonomy approach: know-why, know-how, know-who, know-when, and know-what (Johannisson, 1991). Consistent with these studies, the scale developers have underlined the need to include not only individual but also social and work-related behaviours in the measurement of university-level entrepreneurship education. Consequently, the authors identified entrepreneurial intention and attitude toward innovation as an employee as key indicators for entrepreneurship education at this level, in addition to skills and knowledge (Moberg et al., 2019).

As can be seen, there are multiple dimensions that entrepreneurship education should focus on to educate entrepreneurial individuals. Therefore, to assess an individual's entrepreneurship education, it is necessary to employ a multidimensional theoretical framework that considers both the knowledge, skills, and attitudes of the individual and behaviours during the entrepreneurial process. Given its multidimensional theoretical content and structure, it is believed that this scale, which was adapted to Turkish, is well-suited to the scope of a typical university-level entrepreneurship education and content validated.

In order to gather evidence supporting the scale's face validity, five sophomores were requested to review the final version of the translation. They addressed the points they didn't understand on paper and online. After that, 3 Turkish language experts were asked to evaluate the articulacy and readability of the scale for the target audience. Finally, the content validity of the scale was theoretically guaranteed by the literature and experts, while face validity was provided by the opinions of both professionals and students.

2.4. Data Collection

Researchers collected their research data from universities in Turkey that were easily accessible. After receiving approval from a university's ethics committee, researchers began data collection. However, they encountered difficulties during the COVID-19 pandemic, and university shutdowns hampered their data collection efforts. As a result, researchers were able to collect 78% of the necessary data (n = 348) before the outbreak in Turkey during the academic year 2018–2019. As a result of analyzing these data, it was determined that the lack of entrepreneurship education experience among the vast majority of participants would negatively impact the validity and reliability of the scale. Therefore, a second round of data collection was performed after the pandemic's effects had subsided. By conducting this second round of data collection, researchers ensured the number of participants was increased by reaching students who have received or are receiving entrepreneurship education at the university through the purposive sampling method. Thus, it was aimed to balance the participants in the data set as much as possible in terms of their entrepreneurship education experience.

2.5. Data Analysis

Since scale adaptation studies involve the adaptation of a tested and adapted model to another language, it is recommended to examine the adaptation of this structure to the Turkish language and the relevant culture with confirmatory factor analysis instead of redetermining the existing structure in the original language with exploratory factor analysis (Seçer, 2018). Confirmatory Factor Analysis (CFA) aims to confirm the structure formed by the relationships predicted theoretically or because of previous analyses (DeVellis, 2003). In this direction, after the process related to the content and face validity of the scale was completed, CFA was conducted with the Amos 24 software, and model fit (construct validity) was examined.

Various methods were used to verify the reliability of the scale. Cronbach’s alpha coefficient was calculated, and reliability was measured by Spearman-Brownan Brown coefficient using split-half method. Cronbach alpha is a useful measure of reliability for multi-item scales and evaluates internal consistency, indicating reliability (Cohen et al., 2007). Split-half method is another way of examining the consistency of responses. For a measurement tool to be reliable, users’ scores should be consistent across items (Creswell, 2012). SPSS 23 software was used for these analyses. Besides composite reliability (construct reliability), a test of internal consistency or internal structure/stability in structural equation modelling (Netemeyer et al., 2003) was calculated.

The analysis process in the study includes a second phase in which group analysis was conducted by some variables. In the literature, there are studies examining the effects of variables such as gender (Vodă & Florea 2019; Tessema Gerba, 2012; Petridou et al., 2009; Wilson et al., 2007; Marques et al., 2018), educational background (Marques et al., 2018), family background on entrepreneurship (Lee et al., 2021; Duval-Couetil et al., 2014; Shinnar et al., 2009; Tessema Gerba, 2012; Peterman & Kennedy, 2003), work experience (Shinnar et al., 2009; Peterman & Kennedy, 2003), entrepreneurship education experience (Lee et al., 2021) on entrepreneurial attitude, desire, intention, and tendency. This study collected data related to these variables and analysed whether entrepreneurship education scores differed according to these variables. Thus, it was aimed to compare the findings with the findings of previous research, and it was examined whether the scale gave consistent results with those revealed by other studies. Therefore, it is thought that these analyses will also support the scale's reliability.

The data collected from the participants were grouped according to the variables (gender, the field of education, volunteering, work experience, experience with self-employment, entrepreneurship education experience, having parents who were born in the same city as where the participants usually live, having parents, or an adult they grew up with, a university degree and having a self-employment acquaintance), and it was evaluated to whether there were significant differences between the groups by independent samples t-test and ANOVA from parametric tests since it was determined that the data showed normal distribution. The effect sizes of the variables with significant differences were also calculated. While the *p*-value reported in group comparisons reports whether there is an effect, it does not reveal the effect size. The *p*-value, which expresses statistical significance, examines whether the findings were by chance. The effect size refers to the size of the differences found (Sullivan & Feinn, 2012). Cohen (1988) categorized the *d* value, which shows the effect size between the two means: small if it is .20, medium if it is .50, and large if it is .80 and above. This value was calculated for analyses using the independent samples t-test in the study. The eta squared value indicating the effect size is calculated when there are more than two groups. Eta squared values are from 1 to 0; 0.01 is evaluated as small, 0.06 as medium, and 0.14 as a large effect (Prajapati et al., 2010). This value was calculated for analysis using ANOVA in the study. SPSS 28 was used to calculate effect sizes.

3. FINDINGS

A confirmatory factor analysis (CFA) was conducted to examine the model fit of EESS and reveal whether the existing model is valid in Turkish. For the scale's reliability, the Cronbach alpha coefficient and Spearman-Brown coefficient with the split-half method were calculated, and the results were analysed.

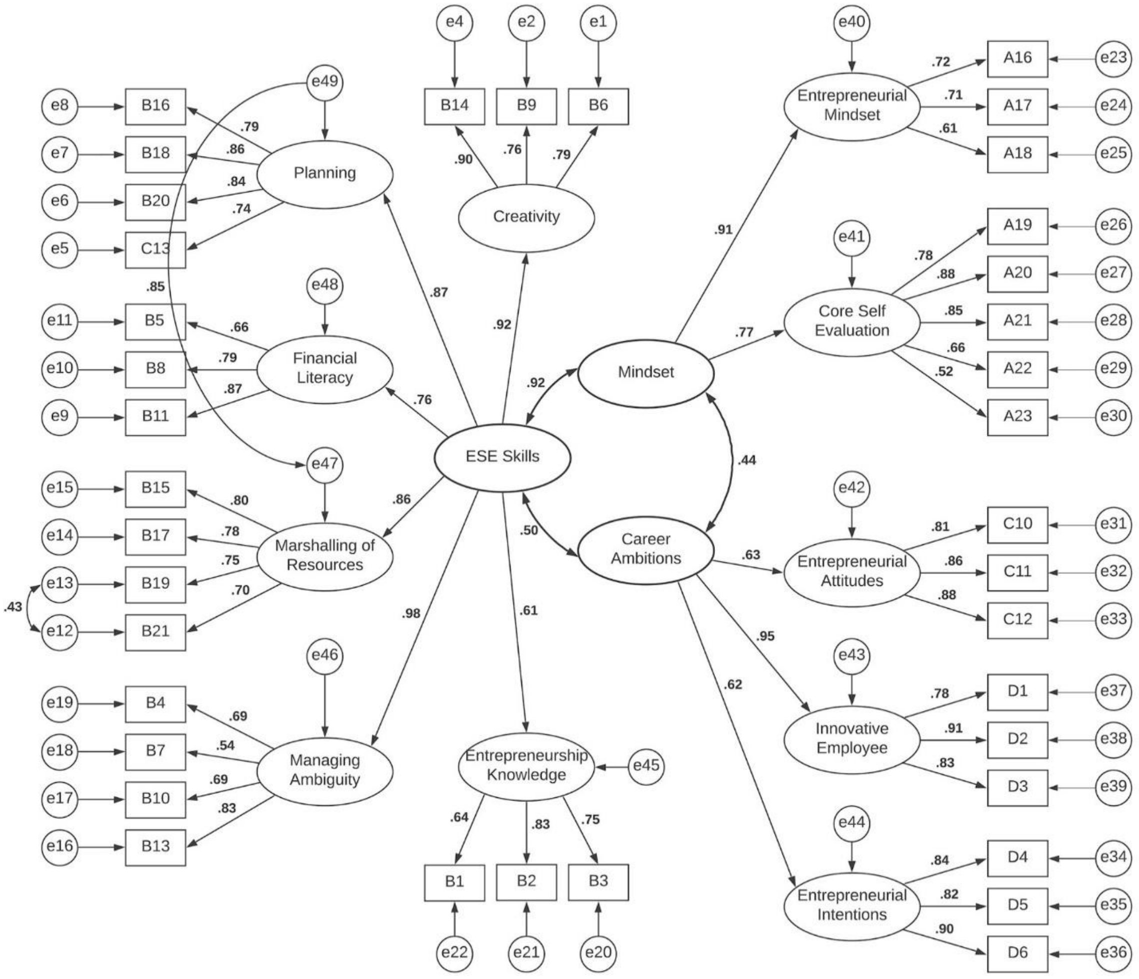
3.1. Confirmatory Factor Analysis

In order to examine the model fit of the Turkish-adapted version of the scale, fit values were calculated and evaluated with CFA. According to the results of the analysis, a very high covariance was found between the error value of the "Entrepreneurial Attitude" sub-dimension under the "Mindset" main dimension and the "Career Ambitions" main dimension. It is stated that the "intention", which is a sub-dimension of career ambitions, depends on the "attitude related to behaviour" (Muofhe & Du Toit, 2011), and in entrepreneurship education, career ambitions are related to entrepreneurship attitude (Dabale & Masese, 2014). Therefore, this sub-dimension was moved from the "Mindset" dimension to the "Career Ambitions" dimension.

After this change, the analysis was repeated, and the modification indices revealed that the covariance between the error value of item 12b (I can identify opportunities for new ways of doing things) and the "Marshalling Resources" sub-dimension was very high. Actively seeking information is one of the key elements that helps entrepreneurs identify various opportunities (Baron, 2006). Yet, as different authors have highlighted, rather than being well planned, such seeks can mostly be performed spontaneously (Ardichvili et al., 2003; Fiet et al., 2004). Kirzner (1985) defined this concept for the first time in the entrepreneurial literature as alertness to changing conditions or reviewing possibilities. This definition demonstrates that people can still find opportunities even if they don't do a detailed examination. This alertness rests, at least in part, on creativity since it helps entrepreneurs identify new solutions. Additionally, it has been proven that, as opposed to only being influenced by creativity (12b belongs to this dimension in the original scale), opportunity recognition is inherently creative (Hansen et al., 2011). Thus, it was seen that item 12b did not have a theoretical direct relationship with the sub-dimension entitled "Marshalling resources". After removing item 12b from the scale, the subsequent analysis revealed a clearer and more accurate representation of the sub-dimension "Marshalling resources". This adjustment ensured that only relevant and meaningful items were included in the assessment. As a result, the revised scale provided a more comprehensive and reliable measure of the construct under investigation. In addition, all modifications to the scale's Turkish adjustment have been discussed and approved via correspondence with the scale's owner. He has actively participated in the decision-making process and given his permission for the modifications.

After the third analysis, a high covariance value was found between the error values of the "Planning" and "Marshalling Resources" sub-dimensions. According to Kickul et al. (2009), these two concepts are the broad stages of creating a new venture that can be nonlinear and iterative. Further studies (Cox et al., 2002; McGee et al., 2009) have indicated that these two concepts are among the key characteristics of ESE that may be quantified. In an experimental study in which entrepreneurship education students evaluated the change in ESE skills, Karlsson and Moberg (2013) found a significant change in both the "Planning" and "Marshalling Resources" dimensions. Marshalling resources involves the willingness to take risks (Jones & English, 2004). Yet, one of the strategies entrepreneurs employ to eliminate uncertainty in risky opportunities and reduce unpredictability in expected results is to create a plan (Forlani & Mullins, 2000). Therefore, as Moberg et al. (2009) stated, these two sub-dimensions are theoretically related, so a modification was made. Similarly, a modification was made for the high correlation between items 19b (*I am able to network*) and 21b (*I am able to establish new contacts*). Entrepreneurs are actively seeking opportunities by making connections with other individuals and organizations (Baron, 2006). For this reason, how actively entrepreneurs use their networks can vary in terms of the number of connections, background, and change in the process (Greve, 1995). That is, these two items are theoretically related to each other. As a result, the model pictured in Figure 1 was validated.

Figure 1. *Structural equation model of EESS, standard values of items, main and sub-dimensions.*



According to Hooper et al. (2008), it is not necessary and unrealistic to include all indexes in the program output in structural equation modelling, but different indexes reflect different aspects of model fit. The Chi-Squared test, RMSEA, CFI, and TLI values included in two index categories mentioned in Hooper et al. (2008) were used. However, the indexes in the third category are not included because no threshold level is recommended for them; therefore, their interpretation has become more difficult. The Chi-Square/df value was 2.939, RMSEA value was .058, CFI .910, TLI .903, IFI .910 from model fit indices. Chi-Square/df values between 2 and 3, RMSEA values between .05 and .08, CFI and TLI values above .90 indicate an acceptable fit (Baumgartner & Homburg, 1996; Crowley & Fan, 1997). These findings show that the scale adapted to Turkish is a valid measurement tool for university students. Table 3 displays the item distribution and factor loadings by main and sub-dimensions for the EESS.

Table 3. *Factor loadings of EESS items.*

	Mindset		ESE (Skills)						Career Ambitions		
Items	Entrepreneurial mindset	Core self-evaluation	Creativity	Planning	Financial literacy	Marshalling of resources	Managing ambiguity	Entrepreneurial Knowledge	Entrepreneurial attitudes	Innovative employee	Entrepreneurial intentions
16a	.723										
17a	.706										
18a	.615										
20a		.879									
21a		.847									
19a		.778									
22a		.661									
23a		.521									
14b			.896								
6b			.787								
9b			.760								
18b				.858							
20b				.836							
16b				.793							
13c				.743							
11b					.866						
8b					.794						
5b					.659						
15b						.798					
17b						.781					
19b						.753					
21b						.704					
13b							.830				
4b							.694				
10b							.690				
7b							.543				
2b								.830			
3b								.753			
1b								.644			
12c									.879		
11c									.861		
10c									.807		
2d										.896	
3d										.843	
1d										.820	
6d											.910
4d											.831
5d											.780

The factor loads of the items belonging to the “Entrepreneurial Mindset” are .723 and .615; the factor loads of the items belonging to the “Core self-evaluation” are .879 and .521; the factor loads of the items belonging to the “Creativity” are .896 and .760; the factor loads of the items belonging to the “Planning” are .858 and .743; the factor loads of the items belonging to the “Financial Literacy” are .866 and .659; the factor loads of the items belonging to the “Marshalling Resources” are . 798 to .704; factor loads of the items belonging to the “Managing Uncertainty” are between .830 and .543; factor loads of the items belonging to the “Entrepreneurship Knowledge” are between .830 and .644; factor loads of the items belonging to the “Entrepreneurial Attitude” are between .879 and .807; factor loads of the items belonging to the “Innovative Employee” are between .896 and .820; factor loads of the items belonging to the “Entrepreneurial Intentions” are between .910 and .780. For the validity of the model, the factor loadings of the items should be above .50 (Farooq, 2016). The fact that the factor loadings of all items are above .50 supports the construct validity of the EESS.

3.2. Reliability Analysis

The reliability of the EESS, whose validity was proven by confirmatory factor analysis and whose last version consisted of thirty-eight items, was examined with Cronbach’s alpha and Composite Reliability coefficients and split-half method. Cronbach’s alpha coefficient was calculated for the overall scale and each sub-dimension. Cronbach’s alpha coefficient was .95, the reliability coefficient for the “Mindset” sub-dimension was .83, .94 for the “Entrepreneurial Skills” sub-dimension, and .88 for the “Career Ambitions” sub-dimension. The composite reliability coefficient was found to be .98. The Spearman-Brown Correlation coefficient was calculated as .86 in the split-half reliability test. In coefficient calculations, values above .70 indicate an acceptable level of reliability (Wilson & Joye, 2017), while between .80 and .90 means highly reliable, and above .90 means very highly reliable (Cohen et al., 2007). Accordingly, it can be said that the EESS is highly reliable.

3.3. Evaluation of Participants’ EESS Scores in Terms of Various Demographic Variables

The mean scores of females and males, those with and without entrepreneurship training, and those with and without self-employment experience according to the sub-dimensions of EESS are presented in Table 4. Accordingly, it is seen that the means differ based on groups. However, to determine whether this difference is significant, the EESS scores of the participants were analysed correlationally in terms of these variables.

Table 4. *EESS mean scores of participants grouped according to gender, entrepreneurship education, and self-employment variables.*

	Gender				Entrepreneurship Education				Self-Employment			
	Female		Male		Ent. Edu.		Non- Ent. Edu.		Exp.		Control	
Factors	x̄	sd	x̄	sd	x̄	sd	x̄	sd	x̄	sd	x̄	sd
Mindset	5.10	.95	5.21	1.06	5.28	1.05	5.06	.96	5.32	1.04	5.05	.96
Entrepreneurial mindset	4.86	1.07	5.14	1.17	5.23	1.11	4.81	1.09	5.32	1.15	4.78	1.05
Core self-evaluation	5.24	1.07	5.25	1.19	5.31	1.18	5.21	1.08	5.32	1.15	5.21	1.1
ESE (Skills)	4.82	.94	5.17	1.04	5.22	1.01	4.79	.95	5.28	1.02	4.78	.93
Creativity	4.92	1.14	5.37	1.22	5.33	1.18	4.95	1.18	5.43	1.21	4.91	1.14
Planning	5.06	1.2	5.34	1.30	5.42	1.22	5.02	1.23	5.48	1.28	5.01	1.19
Financial literacy	3.62	1.30	4.44	1.39	4.37	1.32	3.67	1.37	4.47	1.35	3.64	1.33
Marshalling of resources	4.99	1.22	5.14	1.37	5.27	1.28	4.92	1.25	5.32	1.32	4.90	1.23
Managing ambiguity	4.60	1.09	5.01	1.16	4.96	1.17	4.63	1.09	5.08	1.09	4.58	1.12
Entrepreneurial Knowledge	5.64	1.04	5.71	1.23	5.98	1.06	5.50	1.11	5.89	1.15	5.56	1.08
Career Ambitions	5.53	1.08	5.86	1.00	5.89	.91	5.52	1.12	6.13	.89	5.40	1.06
Entrepreneurial attitudes	5.86	1.27	5.92	1.12	6.05	1.06	5.79	1.29	6.16	1.1	5.74	1.25
Innovative employee	6.01	1.07	6.18	.89	6.24	.85	5.98	1.08	6.36	.77	5.92	1.08
Entrepreneurial intentions	4.72	1.61	5.47	1.55	5.37	1.48	4.78	1.67	5.88	1.42	4.55	1.55

During the research process, data were collected on the variables of gender, being born in the city where the parents usually live, field of education, volunteering, work experience, having a parent or an adult whom you grew up with a university degree, having someone self-employed, entrepreneurship education experience, and experience with self-employment. In order to conduct the necessary analyses, it was examined whether the student scores showed a normal distribution in terms of these variables, and the analysis methods were decided. Skewness and kurtosis values of the data are among the methods used to examine the normal distribution of a data set (Morgan & Driego, 1998). Table 5 shows the kurtosis and skewness values of the scores according to the variables considered in the study.

Tabachnick and Fidell (2013) state that data sets with kurtosis and skewness values between -1.5 and +1.5 can be normally distributed. When the values in Table 5 are examined, it is seen that the kurtosis and skewness values of the participants' entrepreneurship education scores for all variables are between the accepted limits. Accordingly, it was decided to use parametric tests in the analyses, t-test for independent samples was used for variables with two subgroups, and ANOVA test was used for variables with more than two subgroups.

Table 5. *Skewness - kurtosis values of students' EESS scores according to demographic variables.*

	Variable	Subgroups	Skewness	Kurtosis
1	Gender	Female	-.222	-.320
		Male	-.498	-.103
2	Having parents who were born in the same city with where they usually live	At least one	-.289	-.279
		None	-.290	-.284
3	Field of education	Science	-.445	-.168
		Health	-.265	-.468
		Social	-.198	-.191
4	Volunteering	Yes	-.076	-.543
		No	-.638	.496
5	Work experience	Yes	-.546	-.065
		No	-.107	-.198
6	Having parents or an adult they grew up with, a university degree	Yes	-.285	-.185
		No	-.296	-.363
7	Having self-employed acquaintance	Yes	-.285	-.333
		No	-.318	-.084
8	Entrepreneurship education experience	Yes	-.607	-.110
		No	-.187	-.119
9	Experience with self-employment	Yes	-.876	1.041
		No	-.072	-.304

3.3.1. *Gender*

Independent Samples T-test was conducted to determine whether the scores obtained from the scale showed a significant difference according to gender.

Table 6. *T-test results of EESS scores by gender.*

Group	N	x̄	S	sd	*t*	*p*
Female	364	191.68	31.396	570	.4.04	.000*
Male	208	202.92	33.052			

*p<.01

As seen in Table 6, there is a significant difference between the EESS scores of males and females, $t(570)=4.04$, $p<.01$. The mean scores of male students (x̄=202.92) are higher than the mean scores of female students (x̄=191.68). The findings obtained show that there is a significant relationship between the mean score and the gender of the participants. The effect size value d is .351, indicating an effect between small and medium size.

3.3.2. *Having parents who were born in the same city with where they usually live*

Participants were asked whether their parents were born in the same city from where the participants usually live. The Independent Sample t-Test was used to analyse whether there was a significant relationship between the participants' EESS scores and whether their parents were born in the same city where they live now.

Table 7. *T-test results of EESS scores by having parents who were born in the same city with where participant usually live.*

Group	N	x̄	S	sd	*t*	*p*
At least one	258	194.69	32.907	570	.718	.473*
None	314	196.65	32.067			

**p*>.01

According to Table 7, there is no significant relationship between the participants' EESS scores and the fact that their parents were born in the city with where the participants usually live (*t*(570)=.718, *p*>.01). Although the mean scores of the participants (x̄ =196.65), none of whose parents were born in the same city with where they live, were higher than the mean entrepreneurship education scores of the participants (x̄ =194.65), at least one of whose parents was born in the same city with where they live, this relationship was not significant.

3.3.3. *Field of education*

Participants were asked about the main field of higher education in which they received their education. The answers were categorized into science, social, and health. ANOVA test was used to determine whether the entrepreneurship education scores of the participants showed a significant difference according to their field of study.

Table 8. *ANOVA results of EESS scores by field of education.*

Source of Variance	Sum of Squares	sd	Mean of Squares	F	*p*	Significant Difference
Intergroup	16686.115	2	8343.057	8.129	.000*	Science - Health Science - Social
Intragroup	583994,843	569	1026.353			
Total	600680.958	571				

**p*<.01

Based on Table 8, there is a significant difference between the EESS scores of the students in terms of field of education, F2-569=8.13, *p*<.01. This finding shows that students' scores vary significantly according to their field of study. To determine whether there was a significant difference between which groups, the Scheffe test was used. It is the only multiple comparison procedure that is consistent with ANOVA results. If there is a significant difference between the groups in the ANOVA results, Scheffe guarantees that at least one of the group comparisons will be equally significant. Also, as ANOVA, it has similar robustness to the assumptions of normality and homogeneity while allowing different sample sizes in each group (Ruxton & Beauchamp, 2008). In the sample of the study, there are 182 participants from the field of science, 110 from the field of health, and 280 from the field of social. According to the results of the Scheffe Test, it was found that the EESS scores of the students studying in the field of science (x̄=203.23, S=31.86) were higher than those of the students studying in the fields of health (x̄=188.83, S=34.32) and social studies (x̄=193.64, S=31.22). There is no significant difference between the scores of students in health and social fields. The eta squared value for the effect size was calculated as .028. This value indicates an effect between small and medium size.

3.3.4. *Volunteering*

In the study, participants were asked whether they had volunteered in a youth organization, a club, or another non-governmental organization. In order to determine whether the EESS scores showed a significant difference according to their volunteering, an Independent Sample t-Test was conducted between the scores that met the necessary assumptions.

Table 9. *T-test results of EESS scores by volunteering.*

Group	N	x̄	S	sd	*t*	*p*
Yes	229	200.36	31.381	570	2.79	.006*
No	343	192.70	32.806			

*p<.01

According to Table 9, students' EESS scores show a significant difference according to whether they have volunteered or not, t(570)=2.79, p<.01 (Table 9). The mean scores of the students who volunteered (x̄ =200.36) were higher than those who did not volunteer (x̄ =192.70). In other words, there is a significant difference between volunteering in a youth organization, a club, or another non-governmental organization and EESS scores. The effect size value d is .238 and indicates a small effect.

3.3.5. *Work experience*

In the study, participants were asked how many years of work experience they had. The participants were divided into two groups: those with part-time or full-time experience and those with no experience, and the data were analysed with an Independent Sample t-Test to determine whether the participants' EESS scores showed a significant difference according to their work experience.

Table 10. *t-test results of EESS scores by work experience.*

Group	N	x̄	S	sd	*t*	*p*
Part-time or full-time work experience	288	201.73	33.292	570	4.5	.000*
No experience	284	189.71	30.423			

*p<.01

Based on Table 10, the EESS scores show a significant difference according to work experience, t(570)=4.5, p<.01, and this difference favors those who work part-time or full-time. The scores of those with part-time or full-time work experience (x̄=201.73) are higher than the entrepreneurship education scores of those without such experience (x̄=189.71). There is a significant relationship between entrepreneurship education and part-time or full-time work experience. The effect size value d is .377, indicating an effect between small and medium size.

3.3.6. *Having parents, or an adult they grew up with a university degree*

Participants were asked whether their parents or any adults they grew up with were university graduates. Whether the EESS scores of the participants showed a significant difference according to whether their parents or an adult they grew up with were university graduates was analysed with the Independent Sample t-Test.

Table 11. *T-test results of EESS scores by having parents or an adult participants grew up with are university graduate.*

Group	N	x̄	S	sd	*t*	*p*
Yes	280	196.17	32.017	570	2.91	.772*
No	292	195.38	32.880			

*p>.01

According to Table 11, the EESS scores do not differ significantly according to whether or not having parents or an adult participant grew up with a university degree, *t*(570)= 2.91, *p*>.01. Based on this result, although the mean scores of the students whose parents or an adult they grew up with have a university degree (x̄=196.17) are higher than the mean scores of the students whose parents or an adult they grew up with (x̄=195.38), this difference is not significant.

3.3.7. *Having a self-employed acquaintance*

Participants were asked whether they had any acquaintance (parent, relative, or friend) who is self-employed. Whether the EESS scores of the participants show a significant difference according to whether they have an acquaintance who is self-employed was evaluated with the Independent Sample t-Test by evaluating the normal distribution of the data.

Table 12. *t-test results of EESS scores by having self-employed acquaintance.*

Group	N	x̄	S	sd	*t*	*p*
Yes	445	196.13	32.450	570	.497	.620*
No	127	194.50	32.475			

**p*>.01

Based on Table 12, there is no significant difference between the EESS scale scores and whether the students have an acquaintance who is self-employed or not, *t*(570)= .497, *p*>.01. There is a very small difference between the mean scores of students who do not have self-employed acquaintance (x̄=194.50) and the mean scores of students who have an acquaintance (x̄=196.13). The results showed that this difference was not significant and that there was no significant relationship between EESS scores and having an acquaintance who owns their own business.

3.3.8. *Entrepreneurship education experience*

In order to determine entrepreneurship education experience, the participants were asked whether they had taken an entrepreneurship course/lesson and whether they had received any extra-curricular activity that focused on entrepreneurship/self-employment. Participants who answered yes to at least one of these two questions were considered those with entrepreneurship education experience, while those who answered no to both questions were considered those without entrepreneurship education experience. Independent Samples T-test was used to determine whether the scores of the two groups on the entrepreneurship scale differed significantly.

Table 13. *T-test results of EESS scores by entrepreneurship education experience.*

Group	N	x̄	S	sd	*t*	*p*
Yes	202	204.89	32.874	570	5.076	.000*
No	370	190.79	31.125			

**p*<.01

According to Table 13, there is a significant difference between the EESS scores of the students who have entrepreneurship education experience and those who did not, *t*(570)=5.076, *p*<.01. The mean score of the participants with entrepreneurship education experience (x̄=204.89) is higher than the mean score of the group with no experience (x̄=190.79). This finding shows that there is a significant relationship between the students' EESS scores and having received/receiving any entrepreneurship education. The effect size value d is calculated as .444, indicating an effect close to medium size.

3.3.9. *Experience with self-employment*

Participants were asked about their experience in starting, running, or setting up their own business to consider their entrepreneurial behaviour in activities outside the curriculum. Participants were grouped into those who had relevant experience in the past or present and those who did not. Independent Samples t-test was used to examine whether the participants' EESS scores differed significantly according to their experience of starting, running, or setting up their own business.

Table 14. *t-test results of EESS scores by experience with self-employment.*

Group	N	x̄	S	sd	*t*	*p*
Yes	191	208.66	30.597	570	7.009	.000*
No	381	189.30	31.420			

*p<.01

According to Table 14, The EESS scores of the students show a significant difference according to the relevant experience, $t(570)=7.009$, $p<.01$. The mean entrepreneurship education scores (x̄=208.66) of the students who started a business in the past, are currently running a business or are trying to start a business are higher than the mean scores (x̄=189.30) of the students who do not have any business venture. This finding shows a significant relationship between the experience of starting/running/setting up their own business and students' entrepreneurship education scores. The effect size value d is calculated as .621, indicating an effect between medium and large size.

4. RESULTS and DISCUSSION

In this study, the adaptation of the tertiary level entrepreneurship education assessment tool developed within the scope of the ASTEE Project into Turkish was conducted. As a result of CFA and reliability tests conducted with 572 participants, an acceptably valid and highly reliable scale consisting of 3 main dimensions, 11 sub-dimensions, and a total of 38 items was obtained (Table 15). The scale, which is a 7-point Likert type in which the participation rates of the items are scored between 1 and 7, can be used to measure and assess the impact of entrepreneurship education students receive in the context of university education.

Table 15. *Entrepreneurship education self-assessment scale factors and number of items.*

Factors	Number of Factor Items	Sub-factors	Number of Sub-factor Items
Mindset	8	Entrepreneurial mindset	3
		Core self-evaluation	5
ESE (Skills)	21	Creativity	3
		Planning	4
		Financial literacy	3
		Marshalling of resources	4
		Managing ambiguity	4
		Entrepreneurial Knowledge	3
Career Ambitions	9	Entrepreneurial attitudes	3
		Innovative employee	3
		Entrepreneurial intentions	3
Total	38	11	38

In the Turkish adaptation of the scale, it was observed that it preserved its original form. However, in the Turkish adaptation, unlike the original, "Entrepreneurial Attitude" was included in a different sub-dimension (Career Ambitions) as a result of the analysis. Liñán et al. (2011) state that entrepreneurial intention relates to personal attitude. As seen in the model of the adapted scale (Figure 1), "Entrepreneurial Attitude" was included in the "Career Ambitions" dimension, which is also the main dimension of "Entrepreneurial Intention". In this case, it can be said that this change is theoretically possible and correct. Another difference in the Turkish scale is that the sub-dimension "Creativity", a sub-dimension of the main dimension "Mindset", consists of 3 items in the Turkish version with the removal of 1 item, while it was 4 items in the original. As a result, the Turkish adaptation of the scale contains 38 items, while the original scale has 39 items.

The study grouped the participants under 9 variables to test how distinctive the scale items were. It was determined that the participants' EESS scores were significantly correlated with their gender, field of education, volunteering, work experience, experience with self-employment, and entrepreneurship education experience, while the scores did not show a significant difference according to having parents who were born in the same city with where the participants usually live, having parents, or an adult they grew up with a university degree and having a self-employment acquaintance. However, it was seen that the effect sizes of the variables that show a significant difference in entrepreneurship education scores differ. Gender, the field of education, and work experience had effect sizes between small and medium; volunteering had a small effect size; the effect size of entrepreneurship education experience is close to medium size; experience with self-employment had an effect size between medium and large. In this case, it can be said that among these variables, experience with self-employment has the greatest influence on entrepreneurship education.

In the study, it was determined that the entrepreneurship education scores of male students were significantly higher than female students, and it was seen that previous studies supported this result. According to the research done by Vodă and Florea (2019) and Tessema Gerba (2012), it has been shown that there are disparities in entrepreneurial intentions between men and women, with mens exhibiting stronger inclinations to engage in entrepreneurial intentions compared to their female counterparts. The study by Petridou et al. (2009) expressed that women feel less confident and less capable of starting entrepreneurial activities than men, even if they have the same education and come from similar backgrounds. Wilson et al. (2007) found that the ESE of female students in middle school, high school, and graduate education was lower than that of male students. Marques et al. (2018) stated that there is a difference between male and female students regarding individual entrepreneurial orientation and that gender affects individual entrepreneurial orientation differently. In the study conducted by Yılmaz and Sünbül (2009), no significant correlation was found between gender and levels of entrepreneurship. However, Büyükyılmaz et al. (2021) reported a significant and partial difference favoring men in terms of gender and perspective on entrepreneurship in their respective studies. As a result, it is seen that there is a situation against women in general in studies related to entrepreneurship. In this context, it can be said that national education programs should plan courses tailored to the specific needs of women, consider their concerns and perceptions about entrepreneurship, and focus on encouraging them (Petridou et al., 2009); in other words, plans and regulations should be made by considering gender differences in entrepreneurship education.

The result that EESS scores differ by the field of education is supported by Marques et al. (2018). In the current study, the scores of the students studying in the field of science were significantly higher than the scores of the students studying in the fields of Health and Social fields, while the scores of the students in the social and health fields did not differ significantly.

Marques et al. (2018) found that entrepreneurship education differs among business, social and human sciences, and engineering majors; the strength of the impact of entrepreneurship education may differ among students depending on the program they have completed. In addition, although entrepreneurship education scores did not differ according to family background in the current study, Marques et al. (2012) explained that having a businessowner in the family has a negative impact on the entrepreneurial intention of individuals, which may be due to the fact that students do not have positive experiences with their family's business activities and see starting and running a business as an undesirable goal. In contrast to this study, Lee et al. (2021) state that the effect of entrepreneurship education on entrepreneurial ambition is strengthened when students have entrepreneurs in their close families. Similarly, Duval-Couetil et al. (2014) state that students with an entrepreneurial family member are more likely to be more interested in starting a business or becoming self-employed. However, According to Shinnar et al. (2009), the presence/absence of entrepreneurs in the family does not have a significant effect on shaping students' behavioural intentions towards entrepreneurship and their employment intention. While this result supports the conclusion of the current research, Tessema Gerba (2012) found no significant difference between the entrepreneurial intentions of students who were exposed to entrepreneurial activities through their families and students who were not exposed to entrepreneurial activities in their families. In this case, it is seen that there is no generalizable result regarding the effect of the relationship of families with entrepreneurship on students.

The study concluded that there is a significant correlation between the employment status of students, whether they have part-time or full-time work, and their scores in entrepreneurship education. The findings of Büyükyılmaz et al. (2021) research align with the outcomes of the present study, indicating that there is a variation in students' perceptions of entrepreneurship based on their prior job experience. Shinnar et al. (2009) state that the student's prior work experience does not have a significant effect on shaping students' behavioural intentions towards entrepreneurship and their employment aspirations. Similarly, Fatoki (2014) found that the entrepreneurial intentions of undergraduate university students in South Africa did not differ significantly according to work experience. Peterman and Kennedy (2003) determined students' entrepreneurial experience with the questions "Have your parents ever started a business?", "Has anyone else you know started a business?", "Have you ever worked for a small or new company?", "Have you ever started a business?". In their study results, they stated that students' exposure to entrepreneurial experiences encouraged their desire to self-employment. Considering the questions, they used to determine entrepreneurial experience, and it is seen that this study supports the results of the current research related to the variables related to these questions, "work experience" and "experience of starting/running their own business" but does not support the results of it related to the variable "having a relative who owns their own business". However, in Atabay and Alamur's (2016) research, it was seen that there is a significant difference between the entrepreneurial tendencies of the students whose mother or father, who provides the family's living, has their own business and those whose mother or father is retired or unemployed. As a result, it is seen that different results have been obtained regarding the effect of work experience on concepts related to entrepreneurship in the studies.

Findings also show that the entrepreneurship education scores of students with entrepreneurship education experience are significantly higher than those without entrepreneurship education. Supporting this result, Vodă and Florea (2019) state that entrepreneurship education prepares young individuals to enter the labour market and provides them with the knowledge, skills, and capacity to take on different challenges. The study by Lee et al. (2021) also supports these results. In the study, it was observed that students who took entrepreneurship courses had higher levels of intense positive emotions towards starting a business than those who did not. In addition, the fact that there was a significant difference between the entrepreneurship scores of the students who had previously taken entrepreneurship education, class, or course and those who had not shown that the scale is distinctive; in other words, it distinguishes students with high entrepreneurship knowledge and skills from those with low entrepreneurship knowledge and skills.

In conclusion, when the research results are evaluated together with the results of related studies in the literature, there are different results are found regarding the effect of demographic variables on entrepreneurship and entrepreneurship education, except gender and entrepreneurship education experience, and there are no generalizable results.

5. LIMITATIONS AND RECOMMENDATIONS

In the study, the Entrepreneurship Education Self-Assessment Scale (EESS), translated and tested in 13 European countries, including Denmark, Sweden, England, France, Italy, and Germany, was adapted from English into Turkish. The researchers encountered some limitations in this process. The first one is about reaching only undergraduate students with entrepreneurship education experience. This group is frequently employed to investigate entrepreneurship because they show a higher propensity toward venture creation than the general population (Liñán & Santos, 2007). Due to the lack of entrepreneurship education policies and strategies at the universities in Turkey, entrepreneurship education is not yet widespread compared to the EU in undergraduate education. For this reason, more participants with entrepreneurship education experience could not be reached in the study. Secondly, the universities that offer entrepreneurship education or stand out with their entrepreneurial university climate are mostly foundation ones that provide education in English, and their number is relatively low in Turkey. Therefore, the participants were mostly from the state universities in Turkey. It is thought that universities may be the subject of future research. Such research results will also reveal the effects of entrepreneurship education in foundation universities and whether there is a difference between both types of higher education institutions.

Based on the results regarding the comparison of entrepreneurship education scores according to certain demographic characteristics, which constitute the second part of the study, it was observed that while there was a significant difference between some variables and entrepreneurship education, some had no effect on the scores. Although it is explained in detail in the discussion section, it is seen that different results that support and do not support each other are obtained in the studies in the literature. In this context, conducting in-depth analyses with qualitative methods in future studies and comparing qualitative results with research results may contribute to the field in terms of generalizability. In addition, examining the entrepreneurship education scores of student groups with different demographic characteristics in quantitative studies can be considered as further research.

Finally, the developed scale aims to meet the needs of researchers, educational policymakers, and, of course, instructors practicing within the domain of entrepreneurship education in Turkish universities. These practitioners can demonstrate the impact of educational designs on students through multiple interrelated variables in accordance with the nature of education rather than a single isolated variable. Using a pre-test post-test quasi-experimental design, practitioners can measure the impact of a course, program, or whole institution of entrepreneurship education in higher education, comparatively examine the impact of entrepreneurship education in different fields or programs or compare mean scores with the EESS data in this study.

The ASTEE scale is limited to the participants' own responses to the items in the relevant sub-dimensions. It may be a good idea to complement these self-assessed measures with course grade scores or additional measures of different entrepreneurial skills, especially creativity.

Researchers or educational policymakers in Turkey can evaluate the mean scores of their sample by comparing them with the participants in this study.

Acknowledgments

Earlier versions of this research have been presented by the researchers as an oral presentation at the 13th International Computer and Instructional Technology Symposium (ICITS 2019) in Kırşehir, Türkiye. 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 authors. **Ethics Committee Number**: Gazi University, 26.06.2019, 2019-195.

Authorship Contribution Statement

Both named authors contributed equally to a collaborative and collective process in conducting this research and in the authorship of this manuscript.

Orcid

Ahmet Çelik https://orcid.org/0000-0001-9511-7516
Ebru Solmaz https://orcid.org/0000-0003-4893-450X

REFERENCES

Aboobaker, N., & D.R. (2020). Human capital and entrepreneurial intentions: Do entrepreneurship education and training provided by universities add value? *On the Horizon*, *28*(2), 73–83. https://doi.org/10.1108/OTH-11-2019-0077

Ahmed, T., Chandran, V.G.R., & Klobas, J. (2017). Specialized entrepreneurship education: Does it really matter? Fresh evidence from Pakistan. *International Journal of Entrepreneurial Behaviour and Research*, *23*(1), 4–19. https://doi.org/10.1108/IJEBR-01-2016-0005

Alum, R.A. (1986). Reviewed work(s): Innovation and entrepreneurship: Practice and principles by Peter F. Drucker. *Public Productivity Review*, *10*(1), 105–109. https://doi.org/10.2307/3380320

Andrijevskaja, J., & Mets, T. (2008). Master program in entrepreneurship and technology management in Estonia. In P. van der Sijde, A. Ridder, G. Blaauw, & C. Diensberg (Eds.). *Teaching Entrepreneurship Cases for Education and Training* (pp. 90–107). Physica-Verlag HD.

Ardichvili, A., Cardozo, R., & Ray, S. (2003). A theory of entrepreneurial opportunity identification and development. *Journal of Business Venturing*, *18*, 105. https://doi.org/10.1016/S0883-9026(01)00068-4

Arkko-Saukkonen, A. (2017). Chapter 10: Connecting businesses, emerging creative talents, and learning environments in an Entrepreneurial University setting the case study of the creative steps. In J. Cunningham, M. Guerrero, & D. Urbano (Eds.), *The World Scientific Reference on Entrepreneurship* (pp. 297-340). https://doi.org/10.1142/9789813220591_0010

ASTEE User Guide. (n.d.). *Assessment tools and indicators for entrepreneurship education*. [User Guide] https://astra.dk/sites/default/files/ASTEE userguide.pdf

Atabay, İ., & Alamur, B. (2016). İşletme ve muhasebe eğitimi alan meslek yüksekokulu öğrencilerinin girişimcilik eğilimlerinin demografik değişkenlere göre incelenmesi [Review of entrepreneurship tendencies of vocational school students studying at business and accounting on the basis of demographic variables]. *Proceedings of the*

Global Business Research Congress, Turkey, 465-474. https://doi.org/10.17261/Pressacademia.2016118668

Bacigalupo, M., Kampylis, P., Punie Y., & Van Den Brande, L. (2016). *EntreComp: The entrepreneurship competence framework. EUR 27939 EN*. Publications Office of the European Union. JRC101581

Balaban, Ö., & Özdemir, Y. (2008). Girişimcilik eğitiminin girişimcilik eğilimi üzerindeki etkisi: Sakarya Üniversitesi İİBF örneği [The Effect of Entrepreneurship Education on Entrepreneurship Inclination: The Case of Sakarya University Faculty of Economics and Administrative Sciences]. *Journal of Entrepreneurship and Development, 3*(2), 133–147.

Bamiatzi, V., Jones, S., Mitchelmore, S., & Nikolopoulos, K. (2015). The role of competencies in shaping the leadership style of female entrepreneurs: The case of North West of England, Yorkshire, and North Wales. *Journal of Small Business Management*, *53*(3), 627–644. https://doi.org/10.1111/jsbm.12173

Baron, R.A. (2006). Opportunity recognition as pattern recognition: How entrepreneurs "connect the dots" to identify new business opportunities. *Academy of Management Perspectives, 20*(1), 104-119. https://doi.org/10.5465/amp.2006.19873412

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, *13*(2), 139-161. https://doi.org/https://doi.org/10.1016/0167-8116(95)00038-0

Béchard, J.P., & Grégoire, D. (2005). Entrepreneurship education research revisited: The case of higher education. *Academy of Management Learning and Education*, *4*(1), 22–43. https://doi.org/10.5465/AMLE.2005.16132536

Bozkurt, Ö.Ç., & Alparslan, A.M. (2013). Girişimcilerde bulunması gereken özellikler ile girişimcilik eğitimi: Girişimci ve öğrenci görüşleri [Characteristics, must be included entrepreneurs and entrepreneurship education: Opinions of entrepreneurs and students]. *Journal of Entrepreneurship and Development*, *8*(1),7-28.

Bulut, Ç. & Aslan, G. (2014). Üniversitelerde girişimcilik eğitimi [Entrepreneurship education in universities]. *The Journal of Social and Economic Research, 14* (27), 1-20. https://doi.org/10.30976/susead.302216

Büyükyılmaz, O., Yıldıran, C., & Ercan, S. (2021). Girişimcilik ve işletme bölümü öğrencilerinin girişimciliğe bakış açılarının yıllar itibariyle karşılaştırılması [Yearly comparison of the opinions of entrepreneurship and business administration department students]. *The journal of Turkish Social Research, 25*(2), 331-344.

Cevher, E. (2016). Yenilikçi girişimciliği geliştirilmesinde girişimcilik eğitiminin önemi: Meslek yüksekokulu öğrencileri üzerine bir araştırma [The importance of entrepreneurship education at the development of innovative entrepreneurship: A research on vocational school students]. *Journal of Social Sciences and Humanities Researches, 17*(37), 1–17.

Chimucheka, T. (2014). Entrepreneurship education in South Africa. *Mediterranean Journal of Social Sciences*, *5*(2), 403–416. https://doi.org/10.5901/MJSS.2014.V5N2P403

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (Sixth Ed.). Routledge.

Cooney, T.M. (2012). Entrepreneurship skills for growth-orientated businesses. In *Denish Business Authority*. http://www.oecd.org/cfe/leed/Cooney_entrepreneurship_skills_HGF.pdf

Cox, L.W., Mueller, S.L., & Moss, S.E. (2002). The impact of entrepreneurship education on entrepreneurial self-efficacy. *International Journal of Entrepreneurship Education, 1*(2), 229–245.

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

Crowley, S.L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Emerging Issues and Methods in Personality Assessment*, *3891*, 285–308. https://doi.org/10.4324/9780203774618-24

Çolakoğlu, H., & Çolakoğlu, T. (2016). Üniversitelerdeki girişimcilik eğitimi ile öz yeterlilik algısı ve girişimcilik potansiyeli ilişkisi üzerine bir saha araştırması [A field study about interaction of entrepreneurship education, self-efficacy perception, entrepreneurship potential in universities]. *Journal of Social Sciences and Humanities Researches, 17*, 70-84.

Dabale, W.P., & Masese, T. (2014). The influence of entrepreneurship education on beliefs, attitudes, and intentions: A cross-sectoral study of Africa university graduates. *European Journal of Business and Social Sciences*, *3*(9), 1–13.

DeVellis, R.F. (2003). *Scale Development: Theory and Applications* (Second Ed.). Sage Publications.

Drucker, P.F. (1982). *The Changing World of the Executive*. Times Books.

Drucker, P.F. (1985a). *Innovation and Entrepreneurship: Practice and Principles*. University of Illinois at Urbana-Champaign’s Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship. https://ssrn.com/abstract=1496169

Drucker, P.F. (1985b). *Innovation and Entrepreneurship* (1st ed.). Harper & Row, Publishers.

Duval-Couetil, N. (2013). Assessing the impact of entrepreneurship education programs: Challenges and approaches. *Journal of Small Business Management*, *51*(3), 394–409. https://doi.org/10.1111/JSBM.12024

Duval-Couetil, N., Gotch, C.M., & Yi, S. (2014). The characteristics and motivations of contemporary entrepreneurship students. *Journal of Education for Business*, *89*(8), 441–449. https://doi.org/10.1080/08832323.2014.933156

Duval-Couetil, N., Reed-Rhoads, T., & Haghighi, S. (2010). Development of an assessment instrument to examine outcomes of entrepreneurship education on engineering students. *Proceedings - Frontiers in Education Conference, FIE, November*. Washington, DC, United States. https://doi.org/10.1109/FIE.2010.5673411

Elert, N., Andersson, F.W., & Wennberg, K. (2015). The impact of entrepreneurship education in high school on long-term entrepreneurial performance. *Journal of Economic Behavior & Organization, 111*, 209–223. https://doi.org/10.1016/J.JEBO.2014.12.020

Ercan, S., & Yıldıran, C. (2021). Bireysel Girişimcilik Yönelimi Ölçeği’nin Türkçe’ye uyarlanması [Turkish adaptation of Individual Entrepreneurship Orientation Scale]. *Journal of Entrepreneurship and Development, 16*(1), 91-105.

European Commission. (2007). *The key competences for lifelong learning – A European framework*. http://ec.europa.eu/education/index_en.html.

European Commission. (2012). *Effects and impact of entrepreneurship programmes in higher education, Internal Market, Industry, Entrepreneurship and SMEs* (Issue March). https://ec.europa.eu/growth/content/effects-and-impact-entrepreneurship-programmes-higher-education-0_en

European Commission. (2016). *A new skills agenda for Europe working together to strengthen human capital, employability and competitiveness*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016DC0381

European Commission. (2019). *Key competences for lifelong learning*. Publications Office. https://doi.org/doi/10.2766/291008

European Commission/EACEA/Eurydice. (2016). *Avrupa’da okulda girişimcilik eğitimi Eurydice raporu [Entrepreneurship education at school in Europe Eurydice report]*. European Union Printing Office.

European Council. (2006). Recommendation of The European Parliament and of The Council of 18 December 2006 on key competences for lifelong learning. *Official Journal of the European Union*, *L 394*, 10–18. http://data.europa.eu/eli/reco/2006/962/oj

European Council. (2013). Entrepreneurship "2020 Action Plan" - Reigniting the entrepreneurial spirit in Europe. *Official Journal of the European Union EN C*, *C 271*, 75–80.

European Council. (2018). Council Recommendation of 22 May 2018 on key competences for lifelong learning Text with EEA relevance. *Official Journal of the European Union*, *C189/1*, 1-13. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AC%3A2018%3A189%3ATOC

Farooq, R. (2016). Role of structural equation modeling in scale development. *Journal of Advances in Management Research*, *13*(1). https://doi.org/10.1108/JAMR-05-2015-0037

Fatoki, O. (2014). The entrepreneurial intention of undergraduate students in South Africa: The influences of entrepreneurship education and previous work experience. *Mediterranean Journal of Social Sciences*, *5*(7), 294–299. https://doi.org/10.5901/mjss.2014.v5n7p294

Fayolle, A. (2008). Entrepreneurship education at a crossroads: Towards a more mature teaching field. *Journal of Enterprising Culture*, *16*(04), 325-337. https://doi.org/10.1142/S0218495808000211

Fayolle, A. (2013). Personal views on the future of entrepreneurship education. *Entrepreneurship & Regional Development*, *25*(7-8), 692-701. https://doi.org/10.1080/08985626.2013.821318

Fayolle, A., & Gailly, B. (2008). From craft to science: Teaching models and learning processes in entrepreneurship education. *Journal of European Industrial Training*, *32*(7), 569–593. https://doi.org/10.1108/03090590810899838

Fayolle, A., & Gailly, B. (2015). The impact of entrepreneurship education on entrepreneurial attitudes and intention: Hysteresis and persistence. *Journal of Small Business Management*, *53*(1), 75–93. https://doi.org/10.1111/JSBM.12065

Fayolle, A., Gailly, B., & Lassas-Clerc, N. (2006). Assessing the impact of entrepreneurship education programmes: A new methodology. *Journal of European Industrial Training*, *30*(9), 701–720. https://doi.org/10.1108/03090590610715022

Fiet, J.O., Clouse, V.G.H., & Norton, W.I., Jr. (2004). Systematic search by repeat entrepreneurs. In J.E. Butler (Ed.), *Opportunity identification and entrepreneurial behavior* (1-27). Information Age Publishing.

Forlani, D., & Mullins, J.W. (2000). Perceived risks and choices in entrepreneurs' new venture decisions. *Journal of Business Venturing*, *15*, 305-322. https://doi.org/10.1016/S0883-9026(98)00017-2

Galloway, L., & Brown, W. (2002). Entrepreneurship education at university: A driver in the creation of high growth firms? *Education + Training*, *44*(8/9), 398–405. https://doi.org/10.1108/00400910210449231

Galvão, A., Marques, C.S., & Ferreira, J. (2019). Evaluation of an entrepreneurship training programme: A proposal for new guidelines. *Education and Training*, *61*(2), 136–152. https://doi.org/10.1108/ET-11-2018-0228

Garavan, T.N., & O'Cinneide, B. (1994). Entrepreneurship education and training programmes: A review and evaluation - Part 1. *Journal of European Industrial Training*, *18*(8), 3–12. https://doi.org/10.1108/03090599410068024

Genç, K.Y. (2019). Eğitim sisteminin girişimcilik davranışı üzerindeki etkisi: Finlandiya örneği [The effect of education system on entrepreneurial behavior: The case of Finland]. *4th International Symposium on Innovative Approaches in Social, Human and Administrative Sciences*, 259–266. https://doi.org/10.36287/setsci.4.8.048

Gibb, A. (2005). The future of entrepreneurship education – Determining the basis for coherent policy and practice? In P. Kyrö & C. Carrier (Eds.), *The Dynamics of Learning Entrepreneurship in a Cross-Cultural University Context.* (pp. 44–68). University of Tampere Research Centre for Vocational and Professional Education.

Gibb, A. (2012). Exploring the synergistic potential in entrepreneurial university development: towards the building of a strategic framework. *Annals of Innovation & Entrepreneurship, 3*(1), 16742. https://doi.org/10.3402/AIE.V3I0.17211

Greve, A. (1995). Networks and entrepreneurship—an analysis of social relations, occupational background, and use of contacts during the establishment process. *Scandinavian journal of management*, *11*(1), 1-24. https://doi.org/10.1016/0956-5221(94)00026-E

Hansen, D.J., Lumpkin, G.T., & Hills, G.E. (2011). A multidimensional examination of a creativity-based opportunity recognition model. *International Journal of Entrepreneurial Behavior & Research*, *17*(5), 515-533. https://doi.org/10.1108/13552551111158835

Hasan, S.M., Khan, E.A., & Nabi, M.N.U. (2017). Entrepreneurial education at university level and entrepreneurship development. *Education + Training*, *59*(7–8), 888–906. https://doi.org/10.1108/ET-01-2016-0020

Heinonen, J., & Poikkijoki, S.-A. (2006). An entrepreneurial-directed approach to entrepreneurship education: mission impossible? *Journal of Management Development, 25*(1), 80-94. https://doi.org/10.1108/02621710610637981

Henderson, R., & Robertson, M. (1999). Who wants to be an entrepreneur? Young adult attitudes to entrepreneurship as a career. *Education + Training*, *41*(5), 236–245. https://doi.org/10.1108/00400919910279973

Henry, C., Hill, F., & Leitch, C. (2005). Entrepreneurship education and training: Can entrepreneurship be taught? Part I. *Education + Training*, *47*(2), 98-111. https://doi.org/10.1108/00400910510586524

Honig, B. (2004). Entrepreneurship Education: Toward a Model of Contingency-Based Business Planning. *Academy of Management Learning & Education*, *3*(3), 258–273. https://doi.org/10.5465/amle.2004.14242112

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Huang-Saad, A., Morton, C., & Libarkin, J. (2016). Unpacking the impact of engineering entrepreneurship education that leverages the Lean LaunchPad Curriculum. *Proceedings – Frontiers in Education Conference, FIE*. Eire, PA, USA. https://doi.org/10.1109/FIE.2016.7757373

Johannisson, B. (1991). University training for entrepreneurship: A Swedish approach. *Entrepreneurship and Regional Development*, *3*(1), 67-82. https://doi.org/10.1080/08985629100000005

Johannisson, B. (2006). University training for entrepreneurship: Swedish approaches. *Entrepreneurship & Regional Development*, *3*(1), 67-82. https://doi.org/10.1080/08985629100000005

Jones, C., & English, J. (2004). A contemporary approach to entrepreneurship education. *Education+ training*, *46*(8/9), 416-423. https://doi.org/10.1108/00400910410569533

Jones, C., & Penaluna, A. (2013). Moving beyond the business plan in enterprise education. *Education and Training*, *55*(8–9), 804–814. https://doi.org/10.1108/ET-06-2013-0077

Karlsson, T., & Moberg, K. (2013). Improving perceived entrepreneurial abilities through education: Exploratory testing of an entrepreneurial self-efficacy scale in a pre-post setting. *The International Journal of Management Education*, *11*(1), 1-11. https://doi.org/10.1016/j.ijme.2012.10.001

Kickul, J., Gundry, L.K., Barbosa, S.D., & Whitcanack, L. (2009). Intuition versus analysis? Testing differential models of cognitive style on entrepreneurial self–efficacy and the new venture creation process. *Entrepreneurship theory and practice*, *33*(2), 439-453.

Kirzner, I.M. (1985). *Discovery and the capitalist process*. University of Chicago Press.

Kyrö, P. (2006). *Entrepreneurship education and finnish society*. Working Papers in Economics, No. 152. Tallinn University.

Lee, Y., Cortes, A.F., & Joo, M. (2021). Entrepreneurship education and founding passion: The moderating role of entrepreneurial family background. *Frontiers in Psychology*, *12*, 743672. https://doi.org/10.3389/fpsyg.2021.743672

Liñán, F., Rodríguez-Cohard, J.C., & Rueda-Cantuche, J.M. (2011). Factors affecting entrepreneurial intention levels: A role for education. *International Entrepreneurship and Management Journal*, *7*(2), 195–218. https://doi.org/10.1007/s11365-010-0154-z

Liñán, F., & Santos, F.J. (2007). Does social capital affect entrepreneurial intentions? *International Advances in Economic Research*, *13*(4), 443-453. https://doi.org/10.1007/s11294-007-9109-8

Liu, H., Kulturel-Konak, S., & Konak, A. (2021). A measurement model of entrepreneurship education effectiveness based on methodological triangulation. *Studies in Educational Evaluation*, *70*, 100987. https://doi.org/10.1016/J.STUEDUC.2021.100987

Malywanga, J., Shi, Y., & Yang, X. (2020). Experiential approaches: Effective pedagogy "for" entrepreneurship in entrepreneurship education. *Open Journal of Social Sciences*, *8*, 311–323. https://doi.org/10.4236/jss.2020.82024.

Marangoz, M., & Taçyu Dolu, Z. (2022). Reasons and Expectations of University students to choose entrepreneurship course. *Journal of Economics Business and Political Researches*, *7*(17), 104–128. https://doi.org/https://doi.org/10.25204/iktisad.947015

Martin, B.C., McNally, J.J., & Kay, M.J. (2013). Examining the formation of human capital in entrepreneurship: A meta-analysis of entrepreneurship education outcomes. *Journal of Business Venturing*, *28*(2), 211-224. https://doi.org/10.1016/J.JBUSVENT.2012.03.002

Marques, C.S.E., Santos, G., Galvão, A., Mascarenhas, C., & Justino, E. (2018). Entrepreneurship education, gender and family background as antecedents on the entrepreneurial orientation of university students. *International Journal of Innovation Science*, *10*(1), 58–70. https://doi.org/10.1108/IJIS-07-2017-0067

Marques, C.S., Ferreira, J.J., Gomes, D.N., & Gouveia Rodrigues, R. (2012). Entrepreneurship education: How psychological, demographic and behavioural factors predict the entrepreneurial intention. *Education + Training*, *54*(8/9), 657-672. https://doi.org/10.1108/00400911211274819

McGee, J.E., Peterson, M., Mueller, S.L., & Sequeira, J.M. (2009). Entrepreneurial self–efficacy: Refining the measure. *Entrepreneurship Theory and Practice*, *33*(4), 965–988. https://doi.org/10.1111/j.1540-6520.2009.00304.x

McMullan, W.E., & Gillin, L.M. (1998). Developing technological start-up entrepreneurs: A case study of a graduate entrepreneurship programme at Swinburne University. *Technovation*, *18*(4), 275–286. https://doi.org/10.1016/S0166-4972(97)00119-3

Mitchelmore, S., & Rowley, J. (2013). Entrepreneurial competencies of women entrepreneurs pursuing business growth. *Journal of Small Business and Enterprise Development*, *20*(1), 125–142. https://doi.org/10.1108/14626001311298448

Moberg, K., Vestergaard, L., Fayolle, A., Redford, D., Cooney, T., Singer, S., Sailer, K., & Filip, D. (2009). *How to assess and evaluate the influence of entrepreneurship education - A report of the ASTEE project with a user guide to the tools*. Young Enterprise. *www.asteeproject.eu*

Morgan, G.A., & Driego, O.V. (1998). *Easy use and interpretation of SPSS for Windows: Answering research questions with statistics* (vol 1). Psychology Press.

Morris, M.H., Webb, J.W., Fu, J., & Singhal, S. (2013). A competency-based perspective on entrepreneurship education: Conceptual and empirical insights. *Journal of Small Business Management*, *51*(3), 352–369. https://doi.org/10.1111/jsbm.12023

Muofhe, N.J., & Du Toit, W.F. (2011). Entrepreneurial education's and entrepreneurial role models' influence on career choice. *SA Journal of Human Resource Management*, *9*(1), 1-15. https://doi.org/10.4102/sajhrm.v9i1.345

Nasr, K. Ben, & Boujelbene, Y. (2014). Assessing the impact of entrepreneurship education. Procedia – Social and Behavioral Sciences, 109, 712-715. https://doi.org/10.1016/J.SBSPRO.2013.12.534

Netemeyer, R.G., Bearden, W.O., & Sharma, S. (2003). *Scaling Procedures: Issues and Applications*. Sage Publications.

OECD/European Union. (2018). *Supporting Entrepreneurship and Innovation in Higher Education in The Netherlands* (OECD Skills Studies). OECD Skills Studies, OECD Publishing. https://doi.org/10.1787/9789264292048-en

Özdemir, P. (2016). Girişimcilik eğitimi ve üniversitelerimiz [Entrepreneurship Education in Turkish Universities]. *Journal of Entrepreneurship and Development, 11*(1), 224-240.

Pazarcık, Y. (2016). Üniversitelerimiz girişimci yetiştirebiliyor mu?: Üniversite öğrencilerinin girişimcilik algısını/eğilimini/özelliklerini ölçen araştırmaların sonuçsal bir değerlendirmesi [Can our universities raise entrepreneurs?: A concluding evaluation of studies measuring university students' entrepreneurship perception/disposition/character istics]. *Journal of Social Sciences and Humanities Researches*, *17*(37 Girişimcilik Özel Sayısı), 140-169. https://dergipark.org.tr/en/pub/sobbiad/issue/36438/413033

Peterman, N.E., & Kennedy, J. (2003). Enterprise Education: Influencing Students' Perceptions of Entrepreneurship. *Entrepreneurship Theory and Practice*, *28(2)*, 129–144. https://doi.org/10.1046/j.1540-6520.2003.00035.x

Petridou, E., Sarri, A., & Kyrgidou, L.P. (2009). Entrepreneurship education in higher educational institutions: the gender dimension. *Gender in Management: An International Journal*, *24*(4), 286–309. https://doi.org/10.1108/17542410910961569

Pittaway, L., & Edwards, C. (2012). Assessment: Examining practice in entrepreneurship education. *Education and Training*, *54*(8), 778-800. https://doi.org/10.1108/00400911211274882

Pittaway, L., Hannon, P., Gibb, A., & Thompson, J. (2009). Assessment practice in enterprise education. *International Journal of Entrepreneurial Behaviour and Research*, *15*(1), 71–93. https://doi.org/10.1108/13552550910934468

Porter, M.E. (1990). *The competitive advantage of nations*. Free Press.

Prajapati, B., Dunne, M., & Armstrong, R. (2010). Sample size estimation and statistical power analyses. *Optometry today*, *16*(7), 10-18.

Rasmussen, E.A., & Sørheim, R. (2006). Action-based entrepreneurship education. *Technovation*, *26*(2), 185–194. https://doi.org/10.1016/j.technovation.2005.06.012

Rideout, E. (2012). *Bounded Rationality and the Supply Side of Entrepreneurship: Evaluating Technology Entrepreneurship Education for Economic Impact* [North Carolina State University]. http://www.papers.ssrn.com/sol3/papers.cfm?abstract_id=2027023

Roman, T., & Maxim, A. (2017). National culture and higher education as pre-determining factors of student entrepreneurship. *Studies in Higher Education*, *42*(6), 993–1014. https://doi.org/10.1080/03075079.2015.1074671

Ruxton, G.D., & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral ecology*, *19*(3), 690-693.

Saeed, S., Muffatto, M., & Yousafzai, S. (2014). A multi-level study of entrepreneurship education among Pakistani University Students. *Entrepreneurship Research Journal*, *4*(3), 1–25. https://doi.org/10.1515/erj-2013-0041

Sakınç, S., & Bursalıoğlu, S.A. (2012). A global change in higher education: Entrepreneurial University Model. *Journal of Higher Education and Science*, *2*(2), 92. https://doi.org/10.5961/jhes.2012.037

Sart, G. (2020). Bireysel girişimcilik eğilimi ölçeğinin geliştirilmesi: Geçerlik ve güvenirlik çalışması [Development of individual entrepreneurship tendency: Validity and reliability study]. *International Journal of Applied Economic and Finance Studies, 1*(5), 58-72.

Schulte, P. (2007). The Entrepreneurial University: A strategy for institutional development. *Higher Education in Europe, 29*(2), 187-191. https://doi.org/10.1080/0379772042000234811

Scott, J., Penaluna, A., & Thompson, J.L. (2016). A critical perspective on learning outcomes and the effectiveness of experiential approaches in entrepreneurship education: Do we innovate or implement? *Education and Training*, *58*(1), 82-93. https://doi.org/10.1108/ET-06-2014-0063

Seçer, İ. (2018). *Psikolojik test geliştirme ve uyarlama süreci SPSS ve Lisrel uygulamaları [Psychological test development and adaptation process SPSS and Lisrel applications]* (2nd ed.). Anı Publishing.

Shamsrizi, M., Pakura, A., Wiechers, J., Pakura, S., & Dauster, D.V. (2021). Digital entrepreneurship for the "decade of action" how entrepreneurs can impact our race towards the sustainable development goals. In *Digital Entrepreneurship impact on business and society* (pp. 303–327). Springer Nature. https://doi.org/10.1007/978-3-030-53914-6

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *The Academy of Management Review*, *25*(1), 217. https://doi.org/10.2307/259271

Sherkat, A., & Chenari, A. (2020). Assessing the effectiveness of entrepreneurship education in the universities of Tehran province based on an entrepreneurial intention model. *Studies in Higher Education*, *47*(1), 97-115. https://doi.org/10.1080/03075079.2020.1732906

Shinnar, R., Pruett, M., & Toney, B. (2009). Entrepreneurship education: Attitudes across campus. *Journal of Education for Business*, *84*(3), 151-159. https://doi.org/10.3200/JOEB.84.3.151-159

Silveyra, G., Herrero, Á., & Pérez, A. (2021). Model of Teachable Entrepreneurship Competencies (M-TEC): Scale development. *International Journal of Management Education*, *19*(1). 1–20. https://doi.org/10.1016/j.ijme.2020.100392

Sullivan, G.M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of-Graduate Medical Education, 4*(3), 279-282. https://doi.org/10.4300%2FJGME-D-12-00156.1

T.C. Kalkınma Bakanlığı. (2014). *10. Kalkınma planı girişimciliğin geliştirilmesi özel ihtisas komisyonu raporu [10th Development plan development of entrepreneurship specialization commission report]*.

Tabachnick, B.G., & Fidell, L.S. (2013). *Using Multivariate Statistics*. Pearson.

Tessema Gerba, D. (2012). Impact of entrepreneurship education on entrepreneurial intentions of business and engineering students in Ethiopia. *African Journal of Economic and Management Studies*, *3*(2), 258–277. https://doi.org/10.1108/20400701211265036

The ENTREDU. (n.d.). *Assessment tools and indicators for entrepreneurship education*. https://portal.opendiscoveryspace.eu/en/content/assessment-tools-and-indicators-entrepreneurship-education-628775

Uygun, M., & Güner, E. (2016). Girişimcilik eğiliminin gelişiminde girişimcilik eğitimin rolü [The role of entrepreneurship education in development of entrepreneurial intent]. *MANAS Journal of Social Studies, 5*(5), 37-57.

Uygun, M., Mete, S., & Güner, E. (2018). Girişimcilik motivasyonu ile girişimcilik eğitimi arasındaki ilişkiler [The Role of Entrepreneurship Education in the Improvement of Entrepreneurship Motivations of Young People]. *Yönetim ve Ekonomi Dergisi, 25*(3), 879–894. https://doi.org/10.18657/YONVEEK.360276

van der Veen, M., & Wakkee, I. (2016). Understanding the entrepreneurial process. In D. Watkins (Ed.), *ARPENT-Annual Review of Progress in Entrepreneurship* (Vol. 2). EFMD.

Varela, R., & Jimenez, J.E. (2001). The effect of entrepreneurship education in the Universities of Cali. *Frontiers of Educational Research: Babson Conference Proceedings, January 2001*, 1–15.

Venkataraman, S. (1997). The distinctive domain of entrepreneurship research. In J. A. Katz (Ed.), *Advances in Entrepreneurship Research: Firm Emergence and Growth* (pp. 119–138). JAI Press.

Vanevenhoven, J., & Liguori, E. (2013). The impact of entrepreneurship education: Introducing the entrepreneurship education project. *Journal of Small Business Management*, *51*(3), 315–328. https://doi.org/10.1111/jsbm.12026

Vodă, A.I., & Florea, N. (2019). Impact of personality traits and entrepreneurship education on entrepreneurial intentions of business and engineering students. *Sustainability, 11*(4):1192. https://doi.org/10.3390/SU11041192

von Graevenitz, G., Harhoff, D., & Weber, R. (2010). The effects of entrepreneurship education. *Journal of Economic Behavior & Organization*, *76*(1), 90-112. https://doi.org/10.1016/J.JEBO.2010.02.015

Wilson, J.H., & Joye, S.W. (2017). *Research methods and statistics: An integrated approach*.

Wilson, F., Kickul, J., & Marlino, D. (2007). Gender, entrepreneurial self–efficacy, and entrepreneurial career intentions: Implications for entrepreneurship education. *Entrepreneurship Theory and Practice*, *31*(3), 387–406. https://doi.org/10.1111/j.1540-6520.2007.00179.x

Yelkikalan, N., Akatay, A., Yıldırım, H.M., Karadeniz, Y., Köse, C., Koncagül, Ö., & Özer, E. (2010). Dünya ve Türkiye üniversitelerinde girişimcilik eğitimi karşılaştırmalı bir analiz [A comparative analysis of entrepreneurship education in World and Turkish Universities]. *KMU Journal of Social and Economic Research*, *12*(19), 51–59.

Yılmaz, E., & Sünbül, A.M. (2009). Üniversite öğrencilerine yönelik girişimcilik ölçeğinin geliştirilmesi [Developing Scale of University Students Entrepreneurship]. *The Journal of Selcuk University Social Sciences Institute*, (21), 195-203.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 672–689

https://doi.org/10.21449/ijate.1255859

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Determining the psychometric properties of middle school statistical thinking testlet-based assessment tool

Lim Hooi Lian[1,*], **Wun Thiam Yew**[1]

[1]School of Educational Studies, University Sains Malaysia, 11800 Minden, Penang, Malaysia

ARTICLE HISTORY

Received: Feb. 24, 2023
Revised: Oct. 16, 2023
Accepted: Oct. 23, 2023

Keywords:
Content validity,
Rasch model,
SOLO model,
Statistical thinking,
Construct validity.

Abstract: The majority of students from elementary to tertiary levels have misunderstandings and challenges acquiring various statistical concepts and skills. However, the existing statistics assessment frameworks challenge practice in a classroom setting. The purpose of this research is to develop and validate a statistical thinking assessment tool involving form one (Grade 7) students' statistical thinking. The SOLO model was applied to develop five testlet tasks. Each testlet task involved four components. This study employed the survey methodology to assess the statistical thinking of 356 form one students. Content validity was determined using the Content Validity Index (CVI). The construct validity was determined using Rasch analysis. The results demonstrated that the instrument for assessing the statistical thinking of the form one students was valid and trustworthy. This finding of the study also revealed new evidence that the instrument allowed the teachers to identify the students' progress effectively based on the hierarchical manner of item levels in the testlet format. The instrument was useful in identifying students' statistical thinking levels. The students' ability to respond appropriately to a task at a particular level reveals their degree of cognitive development. Testlet task was also easy to diagnose the strengths and weaknesses in learning statistics topics.

1. INTRODUCTION

Quantitative information is everywhere. Everyone depends on statistical information to make interpretations and decisions. For instance, the outbreak of the coronavirus (COVID-19) pandemic is one of the obvious examples that require government, authorities, and citizens to make informed decisions based on the latest statistics data (Bilgin, Bulger & Fung, 2020). In the current world dominated by quantitative information, it is impossible to avoid tables, charts, raw scores, central tendency, and dispersion values. Consequently, the study of statistics offers a crucial tool for educating the population to respond wisely and logically to quantitative information in their environment.

All citizens should develop this important skill as a part of their daily lives. As stated by Kerka (1995), in order to become rational, creative, and dynamic citizens in today's society, we should equip ourselves with the ability to make use of quantitative information. Consequently, statistics have emerged as one of the core subjects included in the mathematics curriculum of most countries. For instance, statistics is a key component of the secondary as well as

*CONTACT: Lim Hooi Lian ✉ hllim@usm.my School of Educational Studies, University Sains Mallaysia, Pulau Pinang, Malaysia

e-ISSN: 2148-7456 /© **IJATE** 2023

primary mathematics curricula in Malaysia. In Malaysia, teaching and learning statistics start at age 8, namely from year two and continuing through year six. Primary students are taught how to organize and group data, read data from charts, construct pictographs, pie charts, and bar charts, by using the fundamental concept of central tendency to interpret data and solve problems involving data handling in everyday life (Malaysia Ministry of Education, 2017; 2018; 2019). In form one of middle school, students learn data representation and data interpretation with the intention of resolving complex routine problems. In form two, the students study and creatively use the concept of central tendency in the context of non-routine problem-solving (Malaysian Ministry of Education, 2018; 2019).

1.1. Review of Related Studies

However, past research has shown that the majority of students from primary school to tertiary level have misunderstandings and difficulties to acquire numerous statistical concepts and skills. These include statistical graph complexity and reading (Arteage et al., 2015), interpretation of box plots (Pierce & Chick, 2013), data representation (Chick & Pierce, 2012; Saidi & Siew, 2019; Ibnatul et al., 2021; Subanji et al., 2021), and descriptive statistics (Chan, et al., 2013), variability (delMas & Liu, 2005; Matthew & Clark, 2007), distribution (Lee & Meletiou-Mavrothesis, 2003), measures of central tendency (Cooper & Shore, 2008; Olani et al., 2011). Malaysian form four students tend to have inferior statistical reasoning abilities, claimed Chan and Zaleha (2014), Krishnan and Noraini (2014). On the other hand, asserted that Malaysian school-level examinations placed a greater emphasis on students' computational proficiency than on their growth as thinkers. As a result, the students unconsciously form a mental habit that highlights statistics' emphasis on mathematics, but rarely gives equal importance to the knowledge development in their statistical learning, which entails higher-order thinking ability. Obviously, this confusing circumstance suggests that the students are having troubles in mastering statistics.

Consequently, the effort to strengthen teaching and learning statistics has been thoroughly examined in a number of research (e.g., Tishkovskaya & Lancaster, 2012; Krishnan & Noraini, 2014; English & Watson, 2015; Setambah et al., 2019). One of them includes the need to reform statistics teaching and learning is about assessing students' learning outcomes, notably statistical reasoning. Consequently, assessment is crucial to the process of learning and teaching. By creating a more suitable teaching and learning environment, the teachers will be better equipped to lead and direct the students in light of the assessment's findings.

On the creation of the assessment framework, several studies have concentrated on statistical thinking. As an illustration, Jones et al. (2000), Mooney (2002), and Watson (2005) developed a framework for determining the level of middle school students' statistical thinking based on the analysis of their responses to the assigned tasks. In contrast, the methodology for assessing the level of statistical reasoning regarding descriptive statistics among high school students was established by Groth (2003), and Chan and Zaleha (2013). Aoyama (2007) used the paradigm developed by Callingham and Watson (2005) to examine how well high school through to university students could comprehend graphs. Aisah et al. (2018) developed an assessment framework based on the updated Bloom Taxonomy to assess the statistical reasoning of engineering students.

Nevertheless, the current assessment methods make it difficult to put theory into reality in a classroom environment. The application of the qualitative approach, particularly in interviews, is extremely time-consuming for the teacher. This assessment approach can only be used to assess a small number of students. Secondly, the implementation primarily relies on the teacher' understanding of the guiding principles of the model, including the SOLO model, to correctly categorize the responses of their students. Subsequently, it is challenging to identify the interconnection of the statistical concepts, particularly when using fundamental statistics skills to produce the new dimensional information (provide the data with an explanatory context or a set of conditions, like making predictions and providing a reason). Furthermore, the existing assessment frameworks have been specially designed for certain curriculum levels in the certain nations. Hence, the newly developed statistics assessment framework and instrument are crucial in establishing the high degree of validity and reliability of the mathematics assessment framework in Malaysia, particularly at the middle school level.

In order to address these issues, this study offered valuable pointers on the development of statistical thinking assessment practice via two-pronged approaches. The assessment framework might first be developed. Then, it is applied to design the hierarchical structure for the statistics tasks based on the descriptors of the framework. The framework can also be easily adopted or adapted to assess other topics of statistics as well. The students' ability to respond appropriately to the task at a particular level reveals their level of cognitive development. Instead of using a qualitative approach, it demonstrates a more complex use of the SOLO model in reverse (testlet task format). It means that the teachers are able to identify their students' strengths and weaknesses easily by referring to the item level that their students managed to solve. It is very convenient and meaningful to be applied in formative assessment. Besides, the overall performance of the topic can be determined effectively as the data are analyzed quantitatively. Thus, summative assessment can also be implemented using the newly developed assessment framework.

Nevertheless, it is more beneficial to practice this testlet task in the combination of formative and summative assessment. Before giving insightful comments to students, teachers can use the testlet task to properly determine the students' strengths and weaknesses. In the meantime, the summative assessment could be utilized to assess the comprehensive performance of the students after the process of teaching and learning for such a topic.

Hence, the purpose of this was to develop a statistical thinking assessment framework and assessment tool involving form one (Grade 7) students' statistical thinking. Second, the reliability and validity of the newly developed instrument were determined to ensure the application was valid in a classroom setting.

1.2. Development of Statistical Thinking Assessment Framework

Jones et al. (2000) developed an assessment framework to characterize the statistical reasoning of elementary and middle school students. The framework identifies four statistical processes, namely: (1) describing data: which involves trying to establish clear and specific information shown visually, identifying graphical conventions, and making important linkages between the display and the original data; (2) managing and minimizing data which involves mental activities on data such as rating, summarizing and grouping, (3) representing data: the construction of visual representations that demonstrate diverse organizational patterns of data; (4) evaluating and interpreting data: the interpretation of statistical results.

Contrarily, Garfield (2002) created an assessment model for statistical reasoning that took into account the five different types of reasoning: procedural, verbal, idiosyncratic, transitional, and integrated process. Students exhibit knowledge of several statistical elements and indicators at the level of idiosyncratic reasoning. Nonetheless, they frequently employed them without completely comprehending them. Thus, their perception of the meaning is frequently wrong. Consequently, students frequently mix them with irrelevant content. At the verbal reasoning level, students demonstrate their verbal comprehension of specific statistical principles, but they are incapable of applying these concepts to actual behavior. In other words, students are able to provide an accurate definition of a concept, but they may not understand it fully. Further, they could not understand how to combine statistical ideas or methods to reach the level of transitional thinking. They could, however, be able to tell them apart with accuracy.

Students frequently demonstrate the capacity to comprehend and adequately grasp the dimensions of statistical ideas or methods at the procedural reasoning level, but they are unable to integrate them properly. If the students are able to fully understand the statistics concepts and confidently decide on the statistics rules and concepts to be applied, it could be determined that the students have reached this level.

In the meantime, Chan and Ismail (2014) developed a tenth-grade assessment framework that incorporated three types of statistical reasoning employing information technology, namely centre, spread, and distribution reasoning. The students understood these three features as whole entities rather than isolated and separated features. The framework was based on the previous framework developed by Mooney's assessment framework. To assess the students' skills in statistical reasoning, a task-based interview was utilized.

The aforementioned assessment frameworks have substantially contributed to the assessment of students' statistical thinking. Although their applicability is confined to qualitative methods of assessment, such as an interview, they appear less applicable in the actual classroom context. Consequently, the purpose of this work is to address this constraint.

1.3. SOLO Model as Reference in Developing Assessment Framework

This assessment framework of this study was devised using the Structure of the Observed Learning Outcome model (SOLO model). Biggs and Collis created this assessment method in 1982. This model has highlighted cumulative cognitive components and latent hierarchy. It reveals that when students respond to a task given, their structure responses can be analyzed and categorized into five levels, from pre-structural to extended abstract levels. These are the characteristics of the hierarchical levels:

a. Prestructural - the students provide their answers without addressing the issue. They lack comprehension of the purpose of the specified task.
b. Unistructural - the students will answer by concentrating on one or a small number of information that direct tangible aspects offered and allotted to them. The information can be accessed via the stem (problem scenario) or the given graphic.
c. Multistructural - the response of the students is to collect more or all the pertinent information provided to acquire the answer. The information may be used as a recipe in which a series of steps are performed sequentially to complete the task. However, they are not integrated by the students.
d. Relational - the students' response entails integrating all parts of the provided information into a cohesive framework or structure. In other words, the information offered is insufficient to tackle the problem immediately. Instead, it must be interconnected carefully to generate an acceptable answer.
e. Extended abstract - the students reply by generalizing the framework or implying a distinct and more abstract context.

The SOLO model and the idea of the testlet were combined to construct an assessment framework and assessment tool for assessing the statistical thinking skills of middle school students. The purpose of this combination was to provide a more user-friendly and practical instrument for assessing students' level of statistical thinking, diagnosing their strengths and shortcomings in relation to this issue.

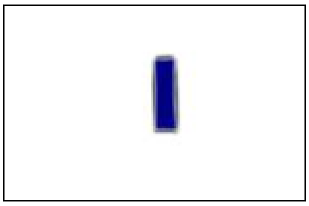
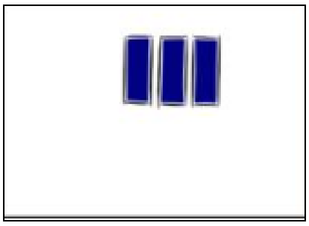
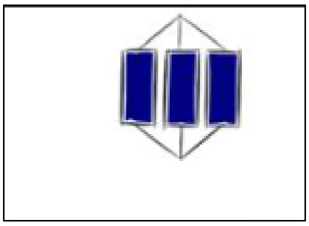
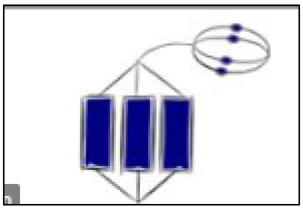
The task's structure consisted of two components. The stem is the initial part of the structure. It describes the situation or issue in the format of paragraphs. The second component is comprised of the four-level items that correspond to the four main levels of the SOLO model. The students' ability to react accurately at a specific level of the item implies that they have attained a particular level of cognitive capacity. The teacher is capable of recognizing the deficiencies of their students depending on the level achieved.

The following criteria were utilized to develop the items for each level:

a. Unistructural – apply directly one observable set of information, source, or material presented in the stem to deliver the answer. When interpreting the information provided, just a single aspect of the information will be considered.
b. Multistructural – employ a more prominent piece of information, or source supplied in the stem to react. At this level, the interpretation of the information may involve more information points without establishing a relationship between them.
c. Relational – use the additional information provided to create a framework by integrating the given information. All accessible information will be integrated at this stage to generate a cohesive meaning and structure.
d. Extended abstract - make deductions and predictions based on knowledge, critical and logical thinking, not simply referring the information in the stem.

Table 1 displays the requirements for the various levels.

Table 1. *Criteria of levels based on SOLO Model.*

Level	The pattern of the response structure	Description
Unistructural		One aspect of the information provided in the task is used to give a response.
Multistructural		Several relevant aspects of the information provided in the task are used to give a response.
Relational		Integrate all the information provided in the task is used to give a response.
Extended abstract		Make hypotheses or predictions for the new situation of the task.

In order to assess middle school students' development in statistical thinking, this study tried to create a thorough and all-encompassing assessment framework. This study specifically focused on two major views in developing an assessment framework: (a) determining the content domains that are essential to the development of statistics thinking in middle school students, and (b) defining and categorizing the levels of statistical thinking across the content domains.

Van de Walle et al. (2014) asserted that in doing statistics, the information analysis process involves three main stages, namely classification, graphical representations, and interpretation of results. Classification refers to deciding how to categorize things. The students normally learn this skill during their early grades. Next, graphical representations and interpretation of findings were the primary focus of this study, as they are taught in middle schools in most nations. In the Malaysian Form One (Grade 7) Secondary School Standard Curriculum, for example, the curriculum requirement of statistics themes comprises the process of gathering, organizing, and representing information, as well as the interpretation of that representation. The following are the key learning standards that encompass the content standards:

a. Construct or convert information representations, such as numerous forms of stem-and-leaf plots, bar charts, dot plots, line graphs, and pie charts.
b. Interpret various information representations, including making inferences or predictions.

In this study, the content domain of the curriculum was utilized to construct the statistical processes for the assessment framework. The cognitive processes of Form one students engaged in the information handling process were represented by four statistical processes: understanding the information provided, calculating, and comparing the value of information, representing the information into various types and making inferences and predictions. On the basis of the four levels of the SOLO model, which are unistructural, relational, multistructural, and extended abstract, the four processes were assessed on four thinking levels. The views of Curcio and Mooney were revised to categorize statistical processes hierarchically, according to the SOLO model. The four levels of information handling implemented by Curcio and Mooney's are as follows:

a. Reading the data: Reading data involves a significantly low cognitive level of data processing. The students merely select the data or facts that are clearly presented in the table or graph, such as the axis labels and titles. At this stage, analysis is unnecessary.
b. Reading between the data: Reading between the data demands a higher level of data handling skill. The interpretation of the data in the table, chart, or graph is included in this level of data handling. It calls for the students' capacity to compare values such as the most significant value, lowest value, minimum points or calculate the quantities such as differences, mean, and range using the fundamental mathematical concepts and skills (such as addition, subtraction, multiplication, and division).
c. Representing data: Graphical representation of data, such as a chart, graph, or dot plot, is known as representing data. This level requires the capacity to organize, interpret and analyze the data. This stage is necessary for displaying the data in a graphical format that reflects the main features of the original data.
d. Reading beyond the data: Reading beyond the data demands students to infer, predict, or make inference using their prior knowledge (or knowledge that they can recall) for the data given. The inference or prediction is neither explicitly nor implicitly stated in the data given.

Table 2. *General statistical thinking assessment framework.*

Category	Unistructural (reading the data)	Multistructural (reading between the data)	Relational (representing data)	Extended abstract (reading beyond the data)
Content Domain of Statistics	The readers simply apply a single data that is explicitly stated in the stem to respond. There is no interpretation at this level.	The readers refer to more or all data in the stem, then: i) apply the basic mathematical concept and skills (e.g., basic mathematics operation) to calculate the total value, differences or percentage; ii) compare the values given.	The readers link all the relevant aspects of data, and incorporate various aspects of his/her statistical thinking so that he/she can convert or represent the data in the appropriate graphical form.	The readers are perceived as having the ability to examine the data from more than one perspective. He/she makes inferences or predictions based on analytic and logical thinking, as well as his/her existing knowledge, not just the data in the stem.

Table 2 displays the basic assessment framework that combines the SOLO model levels and the viewpoints of Curcio and Mooney to classify the hierarchy of statistical thinking to be assessed. Unistructural, multistructural, relational, and extended abstract are the four major levels. At the unistructural level, the student responds by applying a single piece of data expressly provided in the stem. At the multistructural level, the student refers to additional data in the stem and employs some elementary mathematical skills in order to react. The student connects all the pertinent data points at the relational level and converts or represents the data in a suitable graphical representation. At the highest level, the students make inference and predictions using logical and analytical thinking based on both the data in the stem and his or her prior knowledge.

Table 3. *Statistical thinking assessment framework.*

Content Domain	Unistructural (reading the data)	Multistructural (reading between the data)	Relational (representing data)	Extended abstract (reading beyond the data)
Line graph	Apply single data in the stem (the number of students enrolled in the year 2016) to give the response.	Apply two data in the stem to find the difference between the two values.	All the data in the stem is analyzed and converted into a line graph	Make a prediction and provide a logical reason based on his/her existing knowledge and the data in the line graph developed.
Bar chart	Apply single data in the stem (the number of students who go to school by car) to give the response.	Apply all data in the stem and the basic mathematics operation to find the value in percentage.	All the data in the stem is analyzed and converted into a bar chart	Make an inference and provide a logical reason based on his/her existing knowledge and the data in the stem.
Pie chart	Apply single data in the stem (the number of Myvi cars sold in 2017) to give the response	Refer to all data in the stem and the basic mathematics operation to find the highest value.	All the data in the stem is analyzed and represented in a pie chart	Make an inference and provide a logical reason based on his/her existing knowledge and the data in the stem.
Dot Plot	Apply single data in the stem (the number of athletes who weigh 46 kilograms) to give the response	Apply all data in the stem and the basic mathematics operation to find the total value.	All the data in the stem is analyzed and converted into a dot plot	Make an inference and provide a logical reason based on his/her existing knowledge and the data in the stem.
Histogram	Apply single data in the stem (the number of students who spend 30-34 hours for their individual study) to give the response	Apply all data in the stem and the basic mathematics operation to find the value in percentage.	All the data in the stem is analyzed and represented in a histogram.	Make an inference and provide a logical reason based on his/her existing knowledge and the data in the stem.

The comprehensive assessment framework describing the features of each level for analyzing the relevant content domains is shown in Table 3. The four item levels are designed in accordance based on the assessment framework for the respective content domains. The capacity of students to accurately respond to a specific level of the items indicates their level of statistical thinking, as shown in Table 3.

2. METHOD

2.1. Research Design

In this study, the survey method was utilized. The purpose of using this method was to obtain authentic data from a large population. Besides, this method provides accurate and appropriate data regarding the characteristics of a particular individual, such as feelings, perceptions, attitudes, and knowledge.

2.2. Sample of Study

According to the Department of Education, there are 3720 (76%) form one students studying in the North-East District and 1200 (24%) form one students studying in the South-West District of the National Secondary Schools, in Penang Island. The desired sample size of form one students is 356, namely 270 form one students from six secondary schools of the North-East District and 86 form one students from two secondary school of the South-West District. The eight schools were randomly selected from a list of schools. To ensure that the results should accurately reflect the population’s average performance, form one students from high, middle, and low-performing classes were included in the sampling from each school. The latest school mathematics test result was referred to identify the students’ mathematics performance.

2.3. Instrumentation

In the development of statistical thinking assessment tool, the standard manner of instrument development stages which was suggested by Miller and Lovler (2018) was applied, namely:

1. Define the constructs and purpose of assessment. Literature review and document analysis were the focus at this stage. The concept and definition of statistical thinking was identified through the document analysis and curriculum content analysis. The purpose of the assessment was determined by the issues detected in the current teaching and learning classroom. Then, an accurate assessment model was searched to assist in the development of the instrument.
2. Develop the test plan and item format – The SOLO model had been applied in developing general statistical thinking framework (refer to Table 2). Then, the item format (refer to Figure 1) and number of items for each content domain was designed.
3. Compose the test item for each content domain based on the statistical thinking assessment framework (refer to Table 3)
4. Pilot and review the test item – The first draft of the test items was reviewed and judged by five experts. The result of the content validity evaluation was quantified using the CVI (Content Validity Index) formula (refer to Table 4). Then, the test was piloted to ensure the language and instructions were clear and easy to understand, as well as to revise the poor items detected.
5. Validate the test and conduct item analysis – After the test was revised. the validation and item analysis processes were gone through based on the data collected from 356 samples.

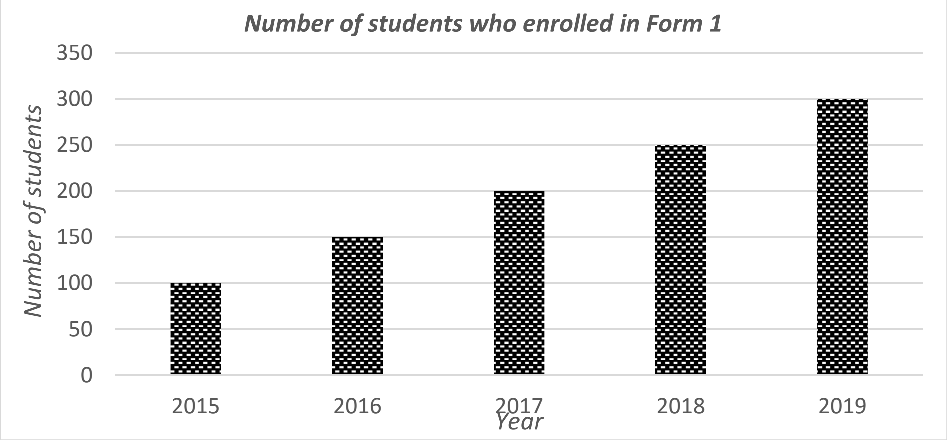
The assessment framework was referred to in creating the testlet task (refer to Table 3). The four items comprising the four levels of the SOLO Model were placed following the stem in the testlet task. Item 1 asked students to refer to a value on the displayed bar chart to produce the right response, which was the number of students enrolled in the 2016 academic year. In item 2, students were required to find the difference between the number of students registered in 2015 and 2019 using the bar chart provided. In item 3, all the values displayed in the bar graph were analyzed and converted into a new graphical form, namely a line graph. In the final item, students were required to make a prediction based on the line graph they created. Afterwards, students were required to produce an explanation based on their logical thinking and prior knowledge, as well as the data in the line graph.

The instrument of this study was composed of five testlets: a line graph, a bar graph, a pie chart, a dot plot, and a histogram. These five testlets reflected the five content domains of the framework. Each testlet included four items. Consequently, there were 20 items in all. All the items were developed in an open-ended format. An example of a line graph testlet created for this study is shown in Figure 1.

The most important procedure after developing the new instrument is to determine its validity. According to AERA (American Educational Research Association) et al. (2014), validity is the degree of evidence supporting the interpretations of the assessment score obtained from the suggested assessment tool. Thus, in this study, the newly developed framework and testlet task required a clear and explicit definition of the domain to be assessed. The validity of the evidence based on the assessment framework and the testlets must be obtained before it can be implemented. In a non-statistical form of validity known as content validity, the content of the test is systematically examined to ensure that it represents a representative sample of the behavior domain being assessed (Anastasi & Urbina, 1997).

Figure 1. *Example of a line graph testlet.*

The number of form one students who registered at SMK Sungai Pasir during a five-year period is depicted in the accompanying bar chart.



a. How many students were enrolled in 2016?
b. Calculate the difference in the number of students enrolled in 2015 and 2019.
c. Convert the bar chart into a line graph.
d. Is it possible to expand your line graph to forecast student enrollment for 2020? Give a reason.

2.4. Data Collection

Each testlet task has four levels of items. Consequently, there were 20 items in all. The items in the five testlet were in an open-ended format. Researchers administered the test to the students after getting the approval of the State Education Department, the participating schools, teachers, and students. The students were given one hour and thirty minutes to answer the 20 items.

It has always been rather arbitrary to choose the appropriate number of experts. According to Zamanzadeh (2015), it was advised that the chance of agreement be sufficiently controlled by at least five experts. In this study, the coverage of all the testlet tasks and the relevance of each item in terms of content validity were assessed by five experts. Three out of the five experts were secondary school teachers. They had been mathematics teachers for middle school students for more than five years. Two more experts were university lecturers in mathematical education. They had been educators of mathematics education courses for more than eight years. Independently, each expert assessed the tasks involved in the assessment. An independent judgement was necessary to guarantee that the experts were not affected by one another, according to AERA et al. (2014). The created assessment tool was revised and improved based on the experts' comments and opinions.

2.5. Data Analysis

The process of validation should involve gathering evidence that provides a sound scientific basis for the assessment score interpretation. The score interpretation included identifying the construct that the assessment tool was supposed to assess (AERA et al., 2014). To quantify the judgement data for content validity, the item-CVI (I-CVI) and scale-level CVI (S-CVI) were chosen. The construct-based validity evidence was also determined for the assessment tool. The Partial Credit Analysis Rasch Model was used to examine the data gathered from 356 form one students. In order to determine whether the assessment tool which were valid and reliable in terms of item dimensionality, item fit, item polarity, reliability, and separation, it was necessary to examine the construct validity.

3. RESULTS

3.1. Content Validity

In calculating CVI, two features must be determined, namely I-CVI and S-CVI. I-CVI are the proportion of the items of the instrument that receives ratings of three or four for their relevance and scope. The proportion of all items that are deemed to have content validity is known as S-CVI.

The significance of each item was rated by experts, often on a 4-point scale in order to determine an item-level CVI (I-CVI) in terms of relevance to the construct or domain assessed. Then, for each item, the I-CVI, was calculated by dividing the number of experts who gave it a score of three or four by the total number of experts. Each item having an I-CVI of 0.78 or above would be regarded as outstanding (Hair et al., 2014; Polit et al., 2007; Price et al., 1995). Researchers should be aware that when the number of panels increases, the probability of chance agreement decreases (Zamanzadeh et al., 2015).

The testlet-based assessment tool consisted of 5 tasks with 20 items. The tasks showed the high content validity of individual items (I-CVI range: 0.80 to 1.00), 13 items (65%) of the items scored 1 while the rest 7 items (35%) scored 0.80, which means all the items were appropriate, and there was no need for revision or elimination. The test generally showed high degree of content validity.

When there are more than two experts, as is most frequently the case, there are a few approaches to compute the scale-level content validity S-CVI. The most common calculation of the S-CVI/Ave is the average I-CVI across items (Polit et al., 2007).

Based on the feedback from five experts, the content validity for scale (S-CVI) was 0.93, which means the testlet tasks have achieved acceptable S-CVI, as shown in Table 4. The designed framework and the assessment tool achieved an acceptable level of content validity. The assessment tool's scale-level content validity and overall content validity index were both scored highly.

Table 4. *Content validity for an Item I-CVI.*

Question	Sub- question	rating 3 or 4	rating 1 or 2	I-CVI	Interpretation
1	(a)	5	0	1*	appropriate
	(b)	5	0	1*	appropriate
	(c)	5	0	1*	appropriate
	(d)	5	0	1*	appropriate
2	(a)	4	1	0.8*	appropriate
	(b)	5	0	1*	appropriate
	(c)	4	1	0.8*	appropriate
	(d)	4	1	0.8*	appropriate
3	(a)	4	1	0.8*	appropriate
	(b)	5	0	1*	appropriate
	(c)	5	0	1*	appropriate
	(d)	5	0	1*	appropriate
4	(a)	4	1	0.8*	appropriate
	(b)	4	1	0.8*	appropriate
	(c)	5	0	1*	appropriate
	(d)	5	0	1*	appropriate
5	(a)	4	1	0.8*	appropriate
	(b)	5	0	1*	appropriate
	(c)	5	0	1*	appropriate
	(d)	5	0	1*	appropriate

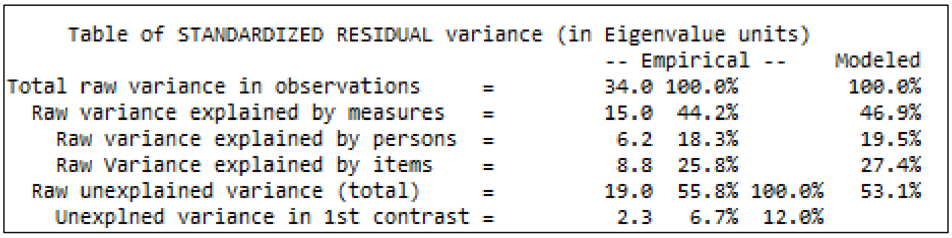
*I-CVI is higher than 79%

3.2. Construct Validity

3.2.1. *Item dimensionality*

The Rasch Model's unidimensionality is one of the most crucial factors in determining the construct validity of the assessment. To ensure that each item is associated with the same latent variable, the item dimensionality was determined. The identical latent variable was assessed, namely statistical thinking in this study (Bond, 2015). Principal Component Analysis (PCA) was analyzed for this component. In Figure 2, the variance explained by the measure was 44.2%. It was close to the model expected, which was 46.9%. The value of the unexplained variance in the first contrast was 6.7%, which was less than 15%, and the eigenvalue for the unexplained variance in the first contrast was 2.3, which is less than 5, confirming that there was no secondary dimension of latent variable that appears in this assessment despite being lower than the expected value and it was also less than 5 (Bond, 2015).

Figure 2. *Standard residual variance.*



```
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                 -- Empirical --    Modeled
Total raw variance in observations     =         34.0 100.0%         100.0%
  Raw variance explained by measures   =         15.0  44.2%          46.9%
    Raw variance explained by persons  =          6.2  18.3%          19.5%
    Raw Variance explained by items    =          8.8  25.8%          27.4%
  Raw unexplained variance (total)     =         19.0  55.8% 100.0%   53.1%
    Unexplned variance in 1st contrast =          2.3   6.7%  12.0%
```

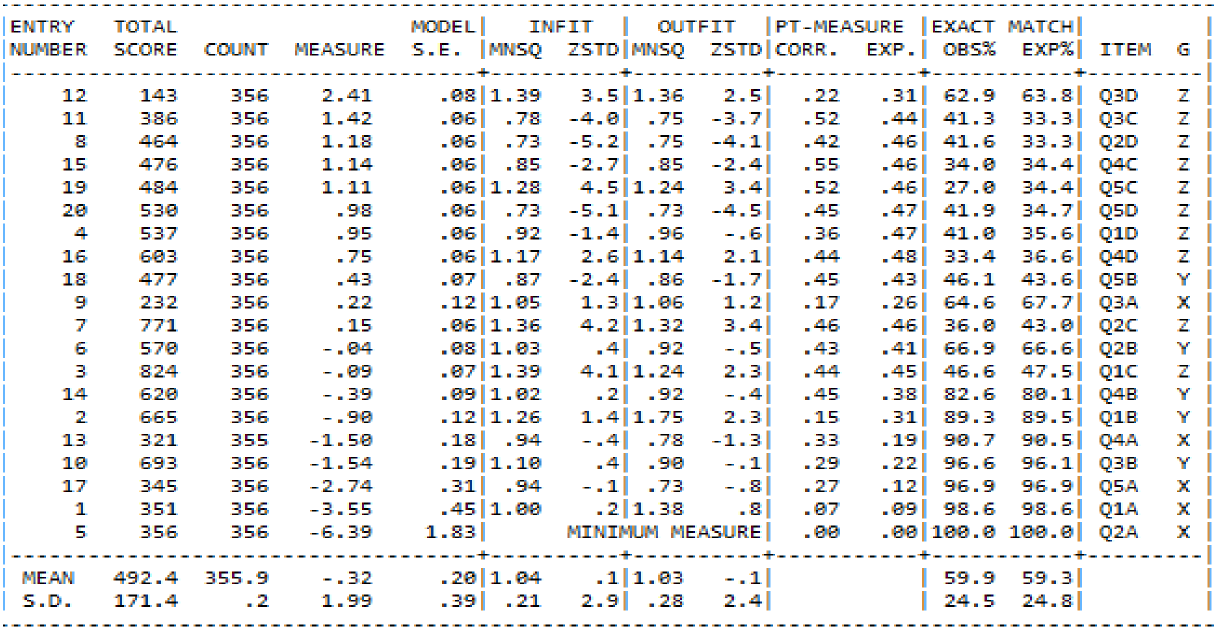
3.2.2. *Item fit*

The item fit has to be determined to ensure all the data collected from the instrument fits the models that contribute to the unidimensionality. The reported fit statistics focused on two aspects: infit and outfit. The acceptable range for the value of infit and outfit was between 0.6 to 1.4 (Lincare, 1994). Figure 3 shows the related result. Every item fell within the acceptable range. Meanwhile, the mean infit MNSQ value for the item was 1.04. The model was able to predict the data too well, as seen by the item's mean outfit MNSQ score of 1.03 According to Linacre (2012), items with an MNSQ value closer to 1.00 was considered a better fit. In conclusion, the finding of the data had the overall fit and was accepted by the Rasch model.

3.2.3. *Item polarity*

The item was able to differentiate between the students' abilities, according to the item's positive PTMEA CORR value. Based on Figure 3, all the items had a positive value of PTMEA CORR, except for Item 5, in which the PTMEA CORR value was 0. It might be that the correlation could not be calculated due to the structure of the data or an extreme item (Bond, 2015). However, almost all the items were able to contribute to the assessment of the latent variable. According to Bond and Fox (2015), the items' good PTMEA CORR values demonstrated their ability to assess the construct that the testlet was supposed to assess and contribute to a high level of construct validity.

Figure 3. *Item statistic: Measure order.*



```
-------------------------------------------------------------------------------------------------
|ENTRY    TOTAL                            MODEL|   INFIT    |  OUTFIT    |PT-MEASURE |EXACT  MATCH|          |
|NUMBER   SCORE   COUNT   MEASURE  S.E.  |MNSQ   ZSTD|MNSQ   ZSTD|CORR.   EXP.| OBS%   EXP%| ITEM   G |
|------------------------------------+----------+----------+-----------+-----------+----------|
|     12    143     356     2.41       .08|1.39    3.5|1.36    2.5|  .22    .31| 62.9   63.8| Q3D    Z |
|     11    386     356     1.42       .06| .78   -4.0| .75   -3.7|  .52    .44| 41.3   33.3| Q3C    Z |
|      8    464     356     1.18       .06| .73   -5.2| .75   -4.1|  .42    .46| 41.6   33.3| Q2D    Z |
|     15    476     356     1.14       .06| .85   -2.7| .85   -2.4|  .55    .46| 34.0   34.4| Q4C    Z |
|     19    484     356     1.11       .06|1.28    4.5|1.24    3.4|  .52    .46| 27.0   34.4| Q5C    Z |
|     20    530     356      .98       .06| .73   -5.1| .73   -4.5|  .45    .47| 41.9   34.7| Q5D    Z |
|      4    537     356      .95       .06| .92   -1.4| .96    -.6|  .36    .47| 41.0   35.6| Q1D    Z |
|     16    603     356      .75       .06|1.17    2.6|1.14    2.1|  .44    .48| 33.4   36.6| Q4D    Z |
|     18    477     356      .43       .07| .87   -2.4| .86   -1.7|  .45    .43| 46.1   43.6| Q5B    Y |
|      9    232     356      .22       .12|1.05    1.3|1.06    1.2|  .17    .26| 64.6   67.7| Q3A    X |
|      7    771     356      .15       .06|1.36    4.2|1.32    3.4|  .46    .46| 36.0   43.0| Q2C    Z |
|      6    570     356     -.04       .08|1.03     .4| .92    -.5|  .43    .41| 66.9   66.6| Q2B    Y |
|      3    824     356     -.09       .07|1.39    4.1|1.24    2.3|  .44    .45| 46.6   47.5| Q1C    Z |
|     14    620     356     -.39       .09|1.02     .2| .92    -.4|  .45    .38| 82.6   80.1| Q4B    Y |
|      2    665     356     -.90       .12|1.26    1.4|1.75    2.3|  .15    .31| 89.3   89.5| Q1B    Y |
|     13    321     355    -1.50       .18| .94    -.4| .78   -1.3|  .33    .19| 90.7   90.5| Q4A    X |
|     10    693     356    -1.54       .19|1.10     .4| .90    -.1|  .29    .22| 96.6   96.1| Q3B    Y |
|     17    345     356    -2.74       .31| .94    -.1| .73    -.8|  .27    .12| 96.9   96.9| Q5A    X |
|      1    351     356    -3.55       .45|1.00     .2|1.38     .8|  .07    .09| 98.6   98.6| Q1A    X |
|      5    356     356    -6.39      1.83|     MINIMUM MEASURE|  .00    .00|100.0  100.0| Q2A    X |
|------------------------------------+----------+----------+-----------+-----------+----------|
| MEAN   492.4   355.9     -.32       .20|1.04     .1|1.03    -.1|           | 59.9   59.3|          |
| S.D.   171.4      .2     1.99       .39| .21    2.9| .28    2.4|           | 24.5   24.8|          |
-------------------------------------------------------------------------------------------------
```

Table 5. *Person- item reliability and separation indices.*

Person/Item	Separation	Reliability
Person	1.37	0.70
Item	9.15	0.99

Based on Table 5, the reliability of a person was 0.70, while the separation of a person was 1.37. The lower value of the person separation value indicated that the limited data available to estimate the students' ability resulted in the lower value of the person reliability value. In contrast, the reliability of the item was 0.99, while the separation of items was 9.15. The value of 0.99 indicated that the item was very consistent. According to Fisher (2007), a reliability value greater than 0.94 is regarded as excellent. A good separation value was more than

2 (Linacre, 2012). It implies that the items might be divided into 9 categories based on the students' levels of ability.

The logit scale was an interval-level measurement scale that was used to both the person's ability and the level of difficulty as shown in Table 6. The four level items in each testlet (except testlet 3: Unistructual to multistructural level, testlet 4 and 5: Relational to extended abstract level) were ordered from the easiest to the most difficult, based on the lowest to the highest value of logit scales shown. It demonstrated that the hierarchy of items were meaaningful. Unistructural level was much easier than the extended abstract level. The relational level was more difficult than multistructural level.

The person locations were plotted to represent that if the students had a 50% probability of correctly answering the item that located at the same point on the logit scale. The students had greater than 50% probability of correctly answering the item with less difficulty, namely the item difficulty level was lower than the students' ability estimate. The bigger gap between the item difficulty and ability estimate, the greater probability of correctly answering the item.

Testlet 1 was an easy item. The four levels of SOLO model were displayed in a hierarchical manner. The majority of students (183) had greater than 50% probability of correctly answering the highest level of item. Meanwhile 152 students had 50% probability of correctly answering the relational level of item.

The four levels of the SOLO model were also displayed in a hierarchical manner in testlet 2. The majority of students (299) had greater than 50% probability of correctly answering the relational level of item. Only 47 students had greater than 50% probability of correctly answering the highest level of item.

The easiest item for testlet 3 was the second level of item, namely multistructural level. Majority of students (270) have greater than 50% probability of correctly answering this item. There was a gap in difficulty level between unistructural level and multistructural level, namely 1.76 logit. Nobody achieved the highest level of this item.

The most difficult item was at the relational level for testlet 4 and 5. There was a small gap between relational and extended abstract levels, namely .39 logits for testlet 4 and .13 logits for testlet 5. However, the majority of students (123) had greater than 50% probability of correctly answering this item.

Table 6. *Statistical thinking level.*

Level/ Testlet	Unistructural (logit scale)	The total number of students who had greater than 50% probability of correctly answering the item	Multistrutural (logit scale)	The total number of students who had greater than 50% probability of correctly answering the item	Relational (logit scale)	The total number of students who had greater than 50% probability of correctly answering the item	Extended abstract (logit scale)	The total number of students who had greater than 50% probability of correctly answering the item
1	-3.55	3	-.90	18	-.09	152	.95	183
2	-6.39	0	-.04	10	.15	299	1.18	47
3	.22	40	-1.54	270	1.42	46	2.41	0
4	-1.50	12	-.39	114	1.14	123	.75	107
5	-2.74	64	.43	109	1.11	123	.98	60

4. DISCUSSION and CONCLUSION

Validity is a unitary concept. It concerns the degree to which various pieces of evidence can be used to support the assessment score (AERA et al., 2014). Thus, evidence about the content of assessments and analyses from the individual response can be gathered to provide sound scientific validity evidence. Content-oriented evidence is the heart of the validation process in educational assessment. It concerns the alignment between content domains to be assessed and the item relevance as well as the content coverage represented by the items (AERA et al., 2014). Content validity index (CVI) is a well-known and widely used formula for quantifying content validity data analysis. There are two features to be determined in applying CVI, namely validity for the item (I-CVI) and content validity for scale (S-CVI). This type of validity offers initial proof of an instrument's construct validity. Additionally, it offers details on the representativeness and clarity of the items. The items might be improved based on the comments and suggestions from a panel of experts (Polit & Beck, 2006; Zamanzadeh et al., 2015).

I-CVI and S-CVI/Ave achieved an acceptable level based on the findings, and the assessment tool's content validity has been attained to a satisfactory level. The development of the testlet task demonstrated high item-content validity for assessing students' statistical thinking. Upon further validation, it is expected that this assessment tool might be used for the development and potential improvement.

The degree of association between the items that conformed to the constructs to be assessed may be determined by analyzing the evidence of validity based on the internal structure (AERA et al., 2014). It might indicate the single dimension of the trait to be assessed. In this study, the unidimensionality, item fit, item polarity, and reliability separation indices were the four primary factors that were required to be justified in order to determine the construct validity based on the principles of the Rasch Model. The newly created assessment tool satisfied the four key requirements stated in the Rasch Model analysis, according to the findings. Besides, Rasch analysis estimated the item difficulty and person ability along a standard scale, namely the logit scale, which provided a more accurate estimation between the difficulty of the item and the ability of a person. In short, the power of Rasch Model analysis enabled the estimation of construct validity regardless of the dependence on person ability (Teh & Lim, 2016). It was an established technique to improve the precision and accuracy in determining psychometric properties for the developed instrument and analyze the respondents' performance characteristics in a more sophisticated manner (Boone, 2016; Teh & Lim, 2016).

As the statistical thinking instrument met the analytical criterion, the results of the content validity and construct validity assessments indicated that it was trustworthy and valid. Therefore, the newly developed assessment framework and assessment tool showed a significant contribution to the new knowledge in statistics assessment in terms of practicality and usefulness. The well-hierarchically organized item level and the quantitative application method of analyzing student's scores proved that the assessment tool could be applied by mathematics teachers either informally (formative assessment) or formally (summative assessment) in their classroom setting.

This study also showed a fresh insight: if students are unable to perform at a higher level, the hierarchical items can assist teachers in identifying the students' deficiencies. This data is essential for teachers to create the most effective remediation strategies for their students. In addition, this testlet task is beneficial for practice in both formative and summative assessments. Before offering helpful feedbacks, teachers will be able to properly assess the students' problems and strengths utilizing the testlet tasks during formative assessment. In the meantime, the summative assessment might be utilized to assess the overall performance of the students after the teaching and learning process for this topic. Both purposes of the assessment can be administered in traditional format (paper-and-pencil test) or the computer mode. In conclusion, the establishment of this assessment framework and assessment tool had significantly contributed to a more effective and efficient assessment of statistical thinking.

Although there are some existing statistical thinking teaching tools developed specifically to the classroom environment such as AutoStat and Code-Driven Tool (Alston-Knox et.al., 2019; Fergusson, 2022), they have some limited fuctions in terms of the content and student level. AutoStat alleviaties the need for coding, emphasizes on statistical models, uses computers and coding to understand the models and algorithms (Alston-Knox et al., 2019). This software is only applicable in higher education level. Meanwhile Code-Driven Tool is a task design framework for developing statistical and computational thinking. The students may need to have some basic knowledge about computer coding in solving the tasks. In sum, if the teachers need to identify the students' strength and weakness in detail during the monitoring and teaching process in statistics topic, these both tools may not be appropriate.

This study provides a more systematic and effective assessment framework as well as a validated assessment tool to assess the middle school students' ability to think critically about the key concept of statistics, namely data handling and representation. Prior to understanding the more advanced concepts of statistics and probability, such as discrete and continuous probability distributions, sampling and estimation (point and interval), and hypothesis testing for population mean and proportion, students must grasp these essential statistical concepts. This assessment framework and assessment tool can be applied in any statistical topic. In addition, it could be utilized as a tool to enhance the students' statistical thinking and to identify their strengths and weaknesses.

Although the statistical thinking assessment tool has been validated, there are some limitations that need to be addressed. First, the assessment tool only included one testlet for each content domain. In future studies, researchers could expand the assessment framework by adding more items for the four levels of the data handling process. Furthermore, the assessment framework was developed to assess only one of the form one mathematics topics, namely statistics. It can be adapted and extended by including the statistics content of form two and form three in the assessment of the students' statistical thinking across the forms.

Acknowledgments

This research was supported by Bridging Grant (304.PGURU.6316212) from Universiti Sains Malaysia, Penang, Malaysia.

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**: Universiti Sains Malaysia, 12345-678.

Authorship Contribution Statement

Lim Hooi Lian: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Wun Thiam Yew**: Methodology and Validation.

Orcid

Lim Hooi Lian https://orcid.org/0000-0002-9089-2262
Wun Thiam Yew https://orcid.org/0000-0002-2714-9636

REFERENCES

Alston-Knox, C.L., Strickland, C.M., Gazos, T., & Mengersen, K.L. (2019). Teaching and Learning in Statistics: Harnessing the power of modern statistical software to improve students statistical reasoning and thinking, *Proceedings of the 5th International Conference on Higher Education Advances (HEAd'19).* http://dx.doi.org/10.4995/HEAd19.2019.9239

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.

Aisah, M.N., MazJamilah, M., Khatijahhusna, A.R., & Safwati, I. (2018). *Developing statistical reasoning and thinking assessment for engineering students: Challenges and new direction.* https://pdfs.semanticscholar.org/c899/c375341f3045d3bd0a5bac0536ce18694fa2.pdf

Anastasi, A., & Urbina, S. （1997）. *Psychological testing (7th ed.).* Prentice Hall.

Aoyama, K., & Stephens, M. (2003). Graph interpretation aspects of statistical literacy: Japanese perspective. *Mathematics Education Research Journal, 15*(3), 3-22.

Arteage, P., Batanero, C., Contreras, J.M., & Canadas, G.R. (2015). Statistical graphs complexity and reading levels: A study with prospective teachers. *Statistique et Enseignement*, *6*(1), 3-23.

Barham, A.I., Ihmeideh, F., Al-Falasi, M., &Alabdallah, A. (2019). Assessment of first grade students' literacy and numeracy levels, and the influence of key factors. *International Journal of Learning, Teaching and Educational Research*, *18*(12), 174-195. https://www.ijlter.org/index.php/ijlter/article/view/1840/pdf

Biggs, J.B., & Collis, K.F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* Academic.

Bilgin, A.A.B., Bulger, D., & Fung, T. (2020). Statistics: Your ticket to anywhere. *Statistics Education Research Journal, 19*(1), 11-20.

Bond, T.G., & Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge.

Callingham, R., & Watson, J.M. (2005). Measuring statistical literacy. *Journal of Applied Measurement*, *6(*1), 29, 19-47.

Chan, S.W., & Zaleha, I. (2012). The role of data technology in developing students' statistical reasoning. *Procedia-Social and Behavioral Sciences*, *46*, 3660-3664.

Chan, S.W., & Zaleha, I. (2014). A technology-based statistical reasoning assessment tool in descriptive statistics for secondary school students. *The Turkish Online Journal of Educational Technology*, *13*(1), 29-46.

Chan, S.W., Zaleha, I., & Bambang, S. (2013). A Rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. *Procedia-Social and Behavioral Sciences 59*[Online], 133 – 139. http://www.sciencedirect.com/science/article/pii/S1877042814028407

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education,18*, 382–393.

delMas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55-82.

English, L., & Watson, J. (2015). Exploring variation in measurement as a foundation for statistical thinking in the elementary school. *International Journal of STEM Education*, *2*(3). https://doi.org/10.1186/s40594-015-0016-x

Fergusson, A-M., G. (2022). *Towards an integration of statistical and computational thinking: Development of a task design framework for introducing code-driven tools through statistical modelling* [Doctor of Philosophy thesis, The University of Auckland]. https://researchspace.auckland.ac.nz/handle/2292/64664

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3). https://doi.org/10.1080/10691898.2002.11910676

Groth, R.E. (2003). *Development of a high school statistical thinking framework* (Unpublished doctoral dissertation), Illinois State University.

Ibnatul, J.F., Adibah, A.L., & Hawa, S.S. (2021). Assessing statistical literacy level of postgraduate education research students in Malaysian research universities. *Turkish Journal of Computer and Mathematics Education*, *12*(5), 1318-1324.

Kerka, S. (1995). *Not just a number: Critical numeracy for adults* (ERIC Digest No. 163, Rep. No.EDO–CE–95–163). Columbus, OH: ERIC Clearinghouse on Adult, Career, and Vocational Education. (ERIC Document Reproduction Service No. ED 385 780).

Krishnan, S., & Idris, N. (2014). Investigating reliability and validity for the construct of inferential statistics. *International Journal of Learning, Teaching and Educational Research*, *4*(1), 51-60.

Linacre, J.M. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.*, 8 (3), p. 370.

Linacre, J.M. (2012). *A user's guide to Winsteps Ministeps Rasch-model computer programs* [version3.74.0]. http://www.winsteps.com/index.htm

Mairing, J.P. (2020). The effect of advance statistics learning integrated Minitab and Excel with teaching teams. *International Journal of Instruction*, *13* (2), 141-150.

Malaysia Ministry of Education (2017). *KSSM Mathematics Form One*. Putrajaya: MOE.

Malaysia Ministry of Education (2018). *KSSM Mathematics Form Two*. Putrajaya: MOE.

Malaysia Ministry of Education (2019). *KSSM Mathematics Form Three*. Putrajaya: MOE.

Matthews, D., & Clark, J. (2007). *Successful students' conceptions of mean, standard deviation and the central limit theorem*. http://www1.hollins.edu/faculty/clarkjm/stats1.pdf

Mooney, E.S. (2002). A framework for characterizing middle school students' statistical. *Mathematical Thinking and Learning*, *4*(1). 23-63.

Olani, A., Hoekstra, R., Harskamp, E., & van der Werf, G. (2011). Statistical reasoning ability, self-efficacy, and value beliefs in a reform-based university statistics course. *Electronic Journal of Research in Educational Psychology*, *9*(1), 49-72.

Pierce, R., & Chick, H. (2012). Workplace statistical literacy for teachers: Interpreting boxplots. *Mathematics Education Research Journal*, *25*, 189-205. http://dx.doi.org/10.1007/s13394-012-0046-3

Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Res Nursing Health*, *29*(5), 489-497.

Polit, D.F., Beck, C.T., & Owen, S.V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, *30*(4), 459–467. https://doi.org/10.1002/nur.20199

Saidi, S.S., & Siew, N.M. (2019). Assessing students' understanding of the measures of central tendency and attitude towards statistics in rural secondary schools. *International Electronic Journal of Mathematics Education*, *14*(1), 73-86. https://doi.org/10.12973/iejme/3968

Setambah, M.A.B., Tajudin, N.M., Yaakob, M.F.M., & Saad, M.I.M. (2019). Adventure learning in basics statistics: Impact on students critical thinking. *International Journal of Instruction*, *12*(3), 151-166. https://doi.org/10.29333/iji.2019.12310a

Subanji., Nusantara, T., Rahmatina, D., & Purnomo, H. (2021). The Statistical Creative Framework in Descriptive Statistics Activities. *International Journal of Instruction*, *14*(2), 591-608. https://doi.org/10.29333/iji.2021.14233a

Tishkovskaya, S., & Lancaster, G.A. (2010). Teaching strategies to promote statistical literacy: Review and implementation. In data and context in statistics education: towards an evidence-based society. *Proceedings of the Eighth International Conference on Teaching Statistics*. International Statistical Institute.

Tishkovskaya, S., & Lancaster, G. (2012). Statistical education in the 21st century: A review of the challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(2), 1–55.

Van de Walle, J.A., Karp, K.S., & Bay-Williams, J.M. (2014). *Elementary and middle school mathematics: Teaching developmentally* (8th ed.). Pearson.

Wild, C.J., & Pfannkuch, M. (1999), Statistical thinking in empirical enquiry, *International Statistical Review*, *67*, 223-265.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: development of an instrument for measuring patient-centred communication. *Journal of Caring Sciences*, *4*(2), 165–178. https://doi.org/10.15171/jcs.2015.017

Zamanzadeh, V., Rassouli, M., Abbaszadeh, A., Majd, H.A., Nikanfar, A., & Ghahramanian, A. (2015). Details of content validity and objectifying it in instrument development. *Nursing Practice Today*, *1*(3), 163–171.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 690–708

https://doi.org/10.21449/ijate.1269977

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

The effect of visual art activities on socialization and stress management of individuals with special needs

Kiymet Bayer[1,*], Seda Liman Turan[2]

[1]Ünye Fehmi Cerrahoğlu Special Education Practice School 3rd Level, Ordu, Türkiye
[2]Trabzon University, Faculty of Education, Department of Fine Arts Education, Trabzon, Türkiye

ARTICLE HISTORY

Received: Mar. 23, 2023
Revised: Sep. 06, 2023
Accepted: Nov. 06, 2023

Keywords:
Special education,
Individual with special needs,
Visual arts activities,
Socialization,
Stress management.

Abstract: This study aims at investigating the impact of visual arts activities on the socialization and stress management of individuals with special needs. This is a qualitative research study that employs "action research" and our data were collected based on the observations of teachers. Over a 20-week period, visual arts activities were carried out with 27 individuals with special needs, including six with autism, seven with Down syndrome, and 14 with moderate to severe intellectual disabilities, who received education at the third level in the "Fehmi Cerrahoğlu Special Education Practice School" in Ordu province during the 2020-2021 and 2021-2022 academic years. The study group included a counselling teacher and 19 special education teachers, who observed the activities and their effects on the socialization levels and stress management of educable individuals with special needs. The data obtained from semi-structured interviews were analyzed using content analysis. Most of the participating teachers agreed that visual arts activities contributed to the socialization and stress management of individuals with special needs, and the study found that these activities played an important role in the inclusion of individuals with special needs in society and led to a decrease in stress symptoms.

1. INTRODUCTION

Socialization is defined as the development of self-perception along with individuals' assimilation of knowledge, skills, values, and tendencies to meet the expectations of society. The relationships between individuals are regulated by the process of socialization. Through socialization, the individual attains both personality and self-identity (Şahan, 2007). Socialization is the sum of the stages that the individual goes through since the moment of birth to obtain a social role. The individual's active participation in social life, expressing oneself and integrating with society, and fulfilling the expected responsibilities enable significant progress in the process of socialization. Individuals who cannot form their personalities and be productive are not socially accepted.

One of the most important problems of individuals with special needs is the inability to communicate and express themselves. It is known that there is prejudice and some negative beliefs against them. The idea that education cannot be done with individuals with special needs and the accompanying pity, ridicule, abstinence, protective behavior... etc. attitudes cause these

*CONTACT: Kıymet BAYER ✉ kymtbayer@gmail.com Ünye Fehmi Cerrahoğlu Special Education Practice School 3rd Level, Ordu, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

individuals to move away from the environment, low self-perception, loss of self-confidence, anxiety, and stress (Baydağ, 2013).

Social relationships also affect the emotional structure of the individuals with special needs. Polloway and Patton (1997) stated in their research that individuals with special needs have difficulty integrating with society due to individual and environmental situations and that these difficulties can be overcome with visual art activities. In their research, Keirstead and Graham (2004) achieved the desired positive goals (peer communication, collaboration, learning by having fun, being able to be integrated into other disciplines... etc.) through art. Grytting (2000), Paksoy (2003), Salderay (2008), Adan (2009), Işık (2014) emphasized that many skills were gained through visual art activities. Salderay (2008) conducted a study with 80 teachers (Visual Art (40) and Special Education (40) working in 55 special education practice schools and teaching visual arts education courses. As a result of this research, it was seen that social and independent living skills improved. Paksoy (2003), in his research, with the research population consisting of 18 students, 8 of whom with special needs at the educable level constituted the cluster and 10 of whom at the normal level constituted the comparison cluster, emphasized the necessity of Visual Arts lessons in the education and development of individuals with special needs aged 8-12. It was stated that Visual Arts activities supported the mental and emotional development of individuals with special needs. As a result of the activities, it was pointed out that individuals expressed themselves, socialized with each other and enjoyed the activities.

Togetherness, creativity, productivity, sharing, and solidarity play an important role in an individual's connection with society. Social relationships, as for any other individual, affect the emotional structure of individuals with special needs. When individuals with special needs experience problems connecting with society, the obstacles they face in expressing themselves cause stress in the individual (Eşsizoğlu et al., 2013), which negatively affects their lives. Cüceloğlu (1994: 321) defines stress as “the effort that individuals spend beyond their physical and psychological limits due to adverse circumstances in the physical and social environment.” Stress may be caused by various factors and is widely recognized as a challenging condition that adversely affects individuals in all aspects. Symptoms related to stress can be mental, emotional, physical, or social. Stress symptoms appear physically, emotionally, mentally and socially. Physical stress symptoms are feelings of anxiety, tension, stomach discomfort, sweating, and imbalance in breathing; emotional stress symptoms are loss of trust, feeling worthless and inadequate, untimely anger attacks, harming oneself and the surroundings; mental stress symptoms are excessive daydreaming, difficulty making decisions, focusing on failure; social stress symptoms are disruption in emotional and social communication, unwillingness to work together, social avoidance and inability to collaborate. Stress management skills are essential for individuals’ emotional well-being, and their lack hinders the socialization of individuals with special needs. Taş (2019) found that the anxiety and stress levels of individuals who had difficulties in conveying their feelings were higher than expected, and positive developments were observed in these individuals where socialization was provided. It was found that individuals who had the opportunity to express themselves through social activities developed confidence and a sense of belonging. Socialization and reducing stressors for individuals with special needs are essential for their inclusion in society. "When individuals participate in social activities, their self-confidence and confidence in their environment increase, they express themselves better, feel belonging to a group, and notice the problems in themselves and their environment earlier" (Taş, 2019:2). A person who socializes and can express himself/herself can also manage stress better.

Visual art activities, which are considered to have multifaceted effects, are also gateways to social relations. Salderay (2010) found that the compatibility of visual arts with social interest and adoption, approval, and appreciation of an individual’s artworks by one’s social environment leads to emotional harmony in the individual. Inner harmony also positively affects external harmony, increasing one's ability to cope with stress. The individual who is validated by the work s/he has done can attain satisfaction. While the individual achieves socialization through the interaction and communication established within the activity, they also realize that they are a member of society. Individuals with special needs fulfill themselves by developing communication skills and taking responsibility.

Thanks to visual art activities, the productivity of the individual with special needs turns into interest. Through visual arts activities, adoption, appreciation, and approval of the individual in the environment cause emotional harmony in the individual. The individual who is approved by the work done can reach satisfaction. The individual, who gains the power to express himself through visual art activities, can initiate the process of communicating with his/her environment with the work he/she produces. Simple art interventions create positive experiences by enabling active creation. Simple art interventions increase self-expression and proficiency by creating behavioral changes and positive bodily perception (Martin et al., 2018). In his research, Eisner (2002) talks about socializing and becoming more qualified through artistic activities. Hetland et al. (2015) stated in his book "Studio Thinking 2: The Real Benefits of Visual Arts Education, Second Edition" that talents emerge through visual arts, and as soon as individuals gain self-confidence by developing their sense of achievement, the socialization process will also be positively affected. The individual who develops communication skills realizes himself by taking responsibility. It also protects physical and mental health. Art education creates mentally healthy individuals with strong communication, especially with visual arts. A sense of duty is formed in individuals who develop manual skills (Aksu, 2011). In addition, as Salderay (2010) states that visual arts are used for therapeutic purposes, that they have a healing power on individuals and patients with special needs, and that they provide improvement in self-perception and spiritual relief. It is possible for individuals with special needs to create new products if they can manage stress. They can gain skills such as sharing with friends, working together, and self-discovery. According to Gündoğdu and Adıgüzel (2016), special needs individuals can enhance their motivation, problem-solving, and decision-making skills, while also learning empathy and improving their capacity to engage in social interactions. "Through art education, it is possible to slow down the behavior of individuals who have excessive aggressive movements and have difficulty controlling themselves" (Erbay, 1995: pp.67). It is believed that individuals with special needs can cope with stress by expressing themselves and interact harmoniously through activities such as visual arts (Taş, 2019).

Meros (1990) concluded in his research that the self-esteem of individuals with special needs and art therapy studies was an effective tool. It is possible to state that visual art activities reduce the stress and anxiety levels of individuals with special needs, which makes them feel better behaviorally. Martin et al. (2018) observed in their research that creative arts therapies and activities in the context of coping with stress significantly reduced stress in participants. In the research conducted by Scott (2017), it has been determined that being involved in artistic activity can reduce clinical and non-clinical stress levels and provide stress management.

The aim of this research is to explore how visual art activities can impact the socialization and stress management of individuals with special needs who are at an educable level. The goal is to use visual art activities in a way that is effective and appropriate for individuals with special needs. To the best of our knowledge, there is no existing research that examines the relationship between visual art activities and the socialization and stress management of individuals with special needs. This study is significant because it can serve as a guide for future academic research and increase our understanding of the importance of visual arts in helping individuals with special needs cope with stress and develop social skills.

1.1. Aim of the Research

The aim of the research is to study the effect of visual art activities regularly applied to individuals with special needs at an educable level on their socialization and control of stress levels.

With this research, it is determined that individuals with special needs socialize with visual arts activities and the contribution of these activities to their stress management; It is aimed to present a sample study to the literature. Thus, we aimed to use visual art activities effectively in the education of individuals with special needs at an adequate level.

1.1.1. *Sub-objectives of the research*

1. Do visual art activities contribute to the socialization of individuals with special needs?
2. Do visual art activities contribute to the stress management of individuals with special needs?

1.2. Rationale and Importance of Research

Visual arts education for individuals with special needs is a systematic educational process in which they work by using visual art applications to achieve their learning goals. The aim of the teacher in visual arts education to create spaces where students enjoy and learn easily. Rodriguez (1984) states that visual arts education can integrate students with each other, plays an active role in acquiring basic skills and values, supports these achievements and other subjects and develops students (Salderay, 2008).

Salderay (2008), Mayer (1999), Polloway and Patton (1997) found that individuals with special needs had difficulty blending into society due to limitations and environment-related situations. They also mentioned that it was possible to overcome these difficulties with visual arts education. According to Salderay (2010), visual arts education develops the socialization skills of the individuals who produce these works, from the process of creating a work to the process of exhibiting it. Visual artworks contribute to the integration of individuals with special needs into society and to the awareness of society about individuals. For this reason, the contribution of visual art activities to the socialization of individuals with special needs is important.

Individuals who need special education gain the ability to produce impressive solutions at the stage of gaining new ideas with different materials and techniques in artistic activities. Visual arts activities enable individuals with special needs to use a different way of expression to express themselves. Individuals with special needs can express their love for their families by drawing a flower or a drawing in a more original way. At the same time, visual art activities can be a communication channel in the expression of the problems they experience and in expressing themselves. In this way, it is possible for individuals with special needs who want to express themselves through visual art activities to gain creative skills.

When the literature was reviewed, there was no research on the socialization and stress management of individuals with special needs through visual art activities. This research is important in terms of being the first in its field in terms of guiding future academic studies, socialization of individuals with special needs, and determining the methods/relationships/ behaviors of coping with stress.

2. METHOD

2.1. Research Design

This is a qualitative research study that employs "action research," one of the qualitative research methods. The research was conducted to "evaluate an application within a predetermined theoretical framework" (Şimşek & Yıldırım, 2005).

The aim of the study is to investigate the effects of regular visual art activities on the socialization levels and stress management of educable individuals with special needs. Due to the difficulty of communication with participants with moderate to severe intellectual disabilities, teachers were enlisted to observe the participants throughout the study. The observations made by the participating teachers were collected and analyzed by the researcher using a semi-structured interview form.

2.2. Study Group

The study group of the research consists of a counselling teacher and 19 special education teachers who work at the Fehmi Cerrahoğlu Special Education Practice School in Ordu province in the 2020-2021 and 2021-2022 academic year and participate in the research voluntarily. Seven of the participants were male and 13 were female. All of the participants received training in special education. 16 teachers are special education graduates, 1 teacher is a graduate of psychological counseling and guidance, and 3 teachers graduated from different fields, attended the 80-hour course and are working as paid teachers in the relevant institution. The ages of the participating students ranged between 14 and 27. The disability types of students with special needs are shown in Table 1. In addition to students who cannot express themselves and have difficulty speaking, their drawing skills were parallel to their intelligence quotient.

Table 1. *Number of individuals with special needs depending on disability types.*

Grade	Number of Individuals with Special Needs Depending on Disability Types		
	Autism	Down Syndrome	Moderate-Severe Mental Disability
9th Grade	2	-	3
10th Grade	-	3	2
11th Grade	2	1	6
12th Grade	2	3	3
Total	6	7	14

2.3. Visual Art Activities

Visual art activities were held in a specially designated area by the visual arts teacher (researcher) outside of class hours, for a total of 27 individuals with special needs, including 6 with autism, 7 with Down syndrome, and 14 with moderate to severe intellectual disabilities, who are in grades 9 to 12. Individual Education Plans (IEPs) were prepared for each student as a result of individual assessments. Practices and activities were organized according to the prepared IEPs. The practitioner who determined the activities was the field expert teacher. The activities were planned and applied for 20 weeks, totaling 50 hours, according to the special education program in effect. These activities included pastel paint, watercolor, acrylic paint, and mixed painting techniques. Painting contents consisted of two-dimensional works such as shape painting, making original art works, mosaic work with pastel paint, watercolor printing works (leaf printing, object printing techniques, etc.), painting from the model with acrylic paint, making an imaginary painting about the selected subject, dripping and painting with mixed technique. In each event, activity selection was carried out by taking into account the individual differences of individuals with medium-to-severe special needs. Choices are left to individuals with special needs and presented to their own preferences, taking into account their own wishes in accordance with their individual characteristics. Material sharing, product evaluation, interpretation, and product sharing took place during the event. In the preparation phase for teaching, pre-teaching evaluation, environment arrangements, attention, motivation, review and initiation of activity were planned and implemented. During the activities, individuals with special needs were observed to socialize with each other and establish a communication network by sharing their work and enjoying themselves while carrying out their artistic works. Due to the fact that individuals with moderate and severe intellectual disability could not express themselves, the data for the study were collected by special education teachers who served as observers. 20 special education teachers, who are also classroom teachers, observed visual art activities and students. The interviews conducted after the teachers' observations sought to determine the impact of these activities on pupils' socialization and stress management.

2.4. Data Collection Tools

The research employed face-to-face interviews as the data collection method. A comprehensive review of the relevant literature was conducted by the researcher and a question pool consisting of 25 questions was created. The prepared questions were reduced to 15 questions in line with the expert opinions and a semi-structured interview form was organized. There were five questions designed to measure the impact of the implemented activities on the socialization of individuals with special needs and ten questions aiming to evaluate the impact of the implemented activities on the stress management of individuals.

The interviews were conducted face-to-face in a quiet and appropriate environment at the school where the research was conducted. The interviews lasted about 10 minutes. Before the interviews, the purpose of the research and the purposes for which the findings of the research would be used were explained to the participants. It was emphasized that participation in the study was voluntary and only volunteer participants were included in the study and data were collected. In addition, in line with the permission of the participants in the study group, the interviews were audio recorded by the researcher.

The content analysis method was used in the analysis of the data. After the data were transferred to the computer environment, the opinions of all participants were analyzed and common themes were determined. These common themes were extracted for each topic and the opinions of the participants were gathered under these themes. The frequency of themes was given in the analyses and the remarkable opinions of the participants about the theme were also used to support the themes.

3. FINDINGS

In this section of the study, the findings obtained from the analysis of research data using the content analysis method are presented under the headings of the effects of visual art activities on the socialization and stress management of individuals with special needs. Under these headings, the questions asked to the participants in the interview were thematized, and 20 different themes were created. These themes were then divided into codes based on the answers provided by the participants and tabulated with code frequencies.

3.1. Teachers' Views on the Socialization of Individuals with Special Needs in Visual Art Activities

The participants were asked 5 questions to determine the effects of applied visual art activities on the socialization of individuals. The findings obtained from the analysis of the data of these questions are presented below.

Question 1: “Were individuals with special needs willing to participate in the activities?”

The findings obtained from the analysis of the answers given by the participants to Question 1 are presented in Table 2.

Table 2. *Findings related to willingness to participate in the activities.*

Theme	Code	Frequency
Willingness to participate in events	Yes, they were.	16
	They came to the lesson in excitement.	2
	They were willing.	1
	They were generally willing but there were some exceptions	4
	1 person with special needs was not very willing, which sometimes presented difficulties.	1
	1 person with special needs was not willing.	1

All the participants stated that individuals were willing to participate in the activities. 11 participants answered that individuals with special needs were willing to participate in the activities; some participants stated that individuals with special needs came to the events running, while others stated that they came to the events excited because it was a course they liked. Participant-10 stated that most of the individuals with special needs liked to paint, so they were very eager. However, 4 participants emphasized that although individuals with special needs were willing to participate in activities in general, there were some exceptions. Participant-6 stated that 90% of the individuals with special needs wanted to participate in the activity, Participant-8 stated that only one individual with special needs did not want to participate, while the others were very happy.

Question 2: “Did they cooperate with their peers, the audience, and the teachers during the activity?”

The findings obtained from the analysis of the answers given by the participants to Question 2 are presented in Table 3.

Table 3. *Findings related to collaboration with peers, audiences and teachers.*

Theme	Code	Frequency
Collaboration with peers, audiences and teachers	Yes, they did.	20
	Cooperation level increased towards the end of the activity.	1
	Cooperation was achieved with the guidance of the teacher.	1
	Cooperation ensued when individuals with special needs became aware of the activity.	1
	The activities increased communication between individuals with special needs, and communication increased cooperation	1

All the participants stated that individuals with special needs cooperated with their peers, the audience, and the teacher. While 11 participants answered the question only as "yes" or "they did", 9 participants provided additional comments. Some of the participants emphasized that the cooperation was there, and it was very nice. Participant-16 stated that cooperation was achieved with the guidance of the teacher. On the other hand, Participant-11 stated that the level of cooperation increased towards the end of the activity. Finally, Participant-17 reported that individuals with special needs became more accustomed to the activity and engaged in more collaboration after understanding the purpose of the activity.

Question 3: “Do you think that participants gained a sense of belonging to a group through visual art activities?”

The findings obtained from the analysis of the answers given by the participants to Question 3 are presented in Table 4.

Table 4. *Findings related to gaining a sense of belonging to a group.*

Theme	Code	Frequency
Gaining a sense of belonging to a group	Yes, I do.	20
	I definitely think so.	3
	Their sense of belonging developed because they willingly participated in the activities.	2
	Since a nice environment was created with the group activity, their sense of belonging developed.	2
	Activity instilled a sense of belonging to the group.	1
	They felt more valuable and communicated more.	1

All the participants reported that the activities gave individuals a sense of belonging to a group. While 11 participants answered this question as "Yes, I do." without commenting, 3 participants expressed their thoughts as "I definitely think so." 6 participants provided additional comments. Participant-10 stated that individuals' sense of belonging developed because they willingly participated in these activities. In addition, Participant-15 pointed out that individuals communicated more with their friends and felt more valuable. Furthermore, Participant-16 and Participant-14 reported that individuals gained a sense of belonging to the group and that they were satisfied with the activities.

Question 4: “Do you think that individuals adapt to social rules and the environment?”

The findings obtained from the analysis of the answers given by the participants to Question 4 are presented in Table 5.

Table 5. *Findings related to compliance with social rules and environment.*

Theme	Code	Frequency
Compliance with social rules and environment	Yes, I do.	20
	They complied with the rules, they were respectful.	1
	Individuals socialized	1
	It was very useful	2
	There was peer communication	2
	They complied with all the rules, including the pandemic rules	1

All the participants stated that individuals adapted to social rules and the environment. While 13 participants answered this question as "Yes, I think" without commenting, 7 participants added explanations to their answers. In these explanations, it was stated that group activities were very useful, and there was peer communication. Participant-10 stated that individuals were respectful and complied with the rules during the activities. Next, Participant-14 pointed out that the activities contributed to the socialization of individuals and that his/her purpose in school is the integration of individuals with society. In addition, Participant-18 emphasized that the events were held on the peak days of the COVID-19 pandemic, everyone used masks and followed other rules. Participants also followed these rules without exception, and in addition, they also complied with the rules of cleanliness and order.

Question 5: “Do you think that individuals' self-confidence has increased?”

The findings obtained from the analysis of the answers given by the participants to Question 5 are presented in Table 6.

Table 6. *Findings related to individuals' self-confidence.*

Theme	Code	Frequency
Their self-confidence	Yes, individuals' self-confidence has increased	20
	Their confidence increased definitely/very much.	3
	Their confidence grew as they succeeded.	2
	Their self-confidence increased.	2
	They got help from each other and their self-confidence increased.	2
	Confidence of individuals increased due to the activity and motivation of the teacher.	1

All of the participants stated that there was an increase in the confidence of individuals. While 6 participants answered this question as "Yes, I think" without commenting, 3 participants answered, "I definitely think" and 1 participant answered "Yes, partially." 10 participants added explanations to their answers. Participant-3 and Participant-4 reported that individuals had an increase in their self-confidence as they succeeded, Participant-8 and Participant-10 stated that individuals got help from each other more frequently and comfortably, and they observed the increase in self-confidence in their classes. In addition, Participant-5 and Participant-17 pointed out that hanging their artwork on the board and applauding each other increased their self-confidence.

3.2. Teachers' Views on the Effects of Visual Art Activities on Stress Management of Individuals with Special Needs

In order to measure the effects of applied visual art activities on individuals' stress management, participants were asked 10 questions. The findings obtained from the analysis of the answers to these questions are presented below.

Question 6: “What was the individual’s level of mastery of the task during the activity?”

The findings obtained from the analysis of the answers given by the participants to Question 6 are presented in Table 7.

Table 7. *Findings related to the individual’s level of mastery of the task during the activity.*

Theme	Code	Frequency
Levels of mastery of their duties	Their mastery of the task was adequate.	20
	Activities were appropriate for their level.	5
	Mastery levels were sufficient with teacher support	4
	They helped each other when needed.	2
	They were resolved to complete the task.	1
	Their level of mastery was surprisingly high.	1

All the participants answered this question by giving positive answers about the individuals’ level of mastery of the task. 17 of the participants evaluated the level of individuals as "adequate" or "definitely sufficient," while 4 participants mentioned the effect of teacher support on their high level of mastery. For example, Participant-4 stated that it was sufficient under the guidance of the practicing teacher. In addition, 5 participants reported that they achieved mastery because the tasks were designed according to the ability of the individuals. Furthermore, Participant-14 also pointed out that individuals did the best they could with the support of their teachers.

Question 7: “Did you observe an individual with special needs who felt under pressure during the activity?”

The findings obtained from the analysis of the answers given by the participants to Question 7 are presented in Table 8.

Table 8. *Findings related to conditions of feeling under pressure.*

Theme	Code	Frequency
Conditions of feeling under pressure	No, I did not.	20
	They came to the class voluntarily and left the class whenever they wanted.	2
	They looked happy.	2
	Appropriate activities were given to individuals	1
	Teacher communication was very good	1

Participants were asked whether they observed any individuals feeling under pressure during the activity. None of the participants answered this question affirmatively. While 14 participants briefly answered as “no” or “I did not,” the six participants gave more detail. Participant-3 and Participant-10 reported that no individual felt under pressure since individuals were in the class on a voluntary basis and could leave whenever they wanted with permission. In addition, Participant-8 and Participant-14 reported that individuals participated in the activities with pleasure and seemed happy during the activities.

Question 8: “Is there an increase or decrease in attention and interest during the activity?

The findings obtained from the analysis of the answers given by the participants to Question 8 are presented in Table 9.

Table 9. *Findings related to* experiencing *an increase/decrease in attention and interest, and their control.*

Theme	Code	Frequency
Experiencing an increase / decrease in attention and interest and their control	There was increased attention and interest	7
	There was increased attention and interest because they liked the lesson and the activity	4
	Some individuals with special needs occasionally experienced low interest and attention.	13
	Distraction was observed in one individual with a physical disability and one special needs individual with autism	2
	Interest occasionally decreased but recovered due to timely and appropriate intervention.	7

Out of 18 participants, 7 of them stated that the attention and interest of individuals with special needs increased. These participants reported that individuals with special needs liked the relevant lesson and that they had a high level of attention and interest because they came to the lesson willingly. 13 participants emphasized that some of the individuals experienced low interest and attention from time to time. They also stated that it was normal for these individuals to lose their interest and attention in a short time due to their conditions. For example, Participant-3 stated that an individual with special needs experienced a lack of attention and interest due to their physical problems, and Participant-9 stated that an individual with autism sometimes lacked attention and interest. However, the 13 participants reported that even if there was a decrease in attention and interest in individuals with special needs, the teacher's timely intervention helped them recover their attention and interest. In addition, Participant-15 stated that sometimes there was distraction, but by attracting attention again with more appropriate activities, the individual's progress in this matter was ensured. Next, Participant-19 also stated that whenever there was a decrease in interest and attention, the teacher integrated the individual into the lesson with different methods, which helped them increase their level of interest.

Question 9: "Did you observe that the visual art activity instigated negative thoughts such as being bored and wanting to finish as soon as possible in individuals with special needs?"

The findings obtained from the analysis of the answers given by the participants to Question 9 are presented in Table 10.

Table 10. *Findings related to negative thoughts such as being bored and wanting to finish the task as soon as possible.*

Theme	Code	Frequency
Negative Thoughts	No, I did not.	20
	Individuals with special needs did the activities willingly and with pleasure.	4
	Individuals with special needs did not want the event to end.	6
	There was no boredom.	1

All the participants stated that none of the individuals with special needs expressed negative thoughts about the course such as finding it boring and long, or a desire to finish it as soon as possible. While 9 participants answered the question briefly as "No, I did not," 11 participants gave their answers by adding additional comments. For example, Participant-11, 12, 13 and 16 reported that, far from being bored, individuals with special needs expressed their desire to do more. Next, Participant-5, 8 and 18 pointed out that individuals with special needs did activities willingly, happily and with pleasure.

Question 10: "Did you observe that the visual art activity instigated a sense of failure in individuals with special needs during the activity?"

The findings obtained from the analysis of the answers given by the participants to Question 10 are presented in Table 11.

Table 11. *Findings related to a sense of failure.*

Theme	Code	Frequency
Sense of failure	No, I did not.	20
	No, I definitely did not.	3
	I observed a sense of achievement.	4
	The activities were appropriate to the level of the individuals.	1

All the participants stated that they did not observe a sense of failure among individuals with special needs during the activities. While 11 participants answered the question briefly as "No, I did not," 9 participants gave their answers by adding additional comments. For example, Participant-5, 13 and 16 stated that they certainly did not observe a sense of failure among individuals with special needs during the activity, on the contrary, they observed a sense of achievement and increased self-confidence. Participant-10 and 14 pointed out that the individuals with special needs performed the tasks very successfully. Finally, Participant-3 emphasized that a person with special needs, who had physical difficulties in holding scissors, had some difficulties in terms of hand coordination at first, but then she also experienced a sense of achievement.

Question 11: "Do you think the individuals with special needs experienced fear, anxiety or restlessness during the activities?"

The findings obtained from the analysis of the answers given by the participants to Question 11 are presented in Table 12.

Table 12. *Findings related to experiencing fear, anxiety or restlessness.*

Theme	Code	Frequency
Experiencing fear, anxiety, or restlessness	No, I do not think so.	20
	Individuals with special needs were happy during the lesson and enjoyed the activities.	2
	The encouragement of the teacher prevented these feelings.	1
	The free and peaceful atmosphere in the classroom prevented these feelings	1

All the participants stated that they did not observe fear, anxiety or restlessness among individuals with special needs during the activities. While 15 participants answered the question briefly as "No, I do not think so," 4 participants made additional comments. Participant-17 stated that individuals with special needs did not experience fear, anxiety and restlessness because they enjoyed activities. Next, Participant-10 emphasized that individuals with special needs did not experience fear, anxiety and restlessness because the teacher comforted them by encouraging them when necessary and by preparing a suitable environment for them. In addition, Participant-16 stated that individuals with special needs did not feel fear, anxiety and restlessness, on the contrary, they felt confident in a free and peaceful environment.

Question 12: "Did you observe that the individual with special needs felt inadequate in making decisions during the activity?"

The findings obtained from the analysis of the answers given by the participants to Question 12 are presented in Table 13.

Table 13. *Findings related to feeling inadequate in making decisions.*

Theme	Code	Frequency
Feeling inadequate in making decisions	No, I did not.	11
	Yes, I observed in some.	9
	The teacher timely intervened when individuals with special needs felt inadequate.	3
	Individuals with special needs, who felt inadequate, learned by asking, and continued the activity	2
	Some experienced difficulties.	1

The Participants were asked whether they observed that individuals with special needs felt inadequate in making decisions during the activities. 11 participants stated that they did not observe that individuals with special needs felt inadequate in making decisions during the activity. On the other hand, 9 participants stated that they observed that some individuals with special needs felt inadequate in decision-making, but the teacher helped them by intervening in such situations in a timely manner. In addition, it was emphasized that individuals with special needs, who felt inadequate in making decisions, improved themselves. Furthermore, Participant-8 stated that individuals with special needs sometimes had difficulties in doing activities, but they learned by asking their friends and teachers, they did not break away from the activity and they continued to do it.

Question 13: "Have you observed that individuals with special needs showing concerns about not being accepted by their peers?"

The findings obtained from the analysis of the answers given by the participants to Question 13 are presented in Table 14.

Table 14. *Findings related to showing concerns about not being accepted by their peers.*

Theme	Code	Frequency
Worry about not being accepted by peers	No, I have not.	20
	Peer communications were on point.	1
	They showed no behavior to indicate such concerns.	1
	They were in tune with their friends.	1
	The teacher created a peaceful and accepting environment.	1

All of the participants stated that they did not observe individuals showing concern that they would not be accepted by their peers. 16 participants answered the question briefly as "No, I have not," while 4 participants made additional comments. Participant-3 emphasized that individuals with special needs did not express any worries about being accepted by their peers, and that peer communication was sufficient and at a high level. In addition, Participant-10 stated that individuals with special needs continued to work in harmony with their friends.

Question 14: "Did you observe emotional instability (constant anxiety, excessive excitement, unwarranted rush, sense of inadequacy, laughing/crying for no reason, etc.) during the activity? If so, what were they?"

The findings obtained from the analysis of the answers given by the participants to Question 14 are presented in Table 15.

Table 15. *Findings related to emotional instability during the activities.*

Theme	Code	Frequency
Emotional instability	No, I did not.	17
	The teacher comforted individuals before the activity.	1
	Problematic behaviors were observed in some individuals with special needs.	3
	The teacher intervened on a few occasions in a timely manner.	1
	Excessive excitement states such as laughing were observed in some individuals with special needs.	1

When the participants were asked whether they observed emotional instability (constant anxiety, excessive excitement, unwarranted rush, sense of inadequacy, laughing/crying for no reason, etc.) during the activity among individuals with special needs, 17 participants stated that they did not, while 3 participants stated that they observed among some individuals with special needs. For example, Participant-5 stated that some individuals with special needs showed excessive excitement behaviors such as laughing. Participant-19, on the other hand, emphasized that sometimes there were situations that can be described as emotional instability, but the teacher included the individuals in the lesson by intervening in a timely manner. Of the 17 participants who stated that they did not observe emotional instability (constant anxiety, excessive excitement, unwarranted rush, sense of inadequacy, laughing/crying for no reason, etc.) among individuals with special needs during the activity, only Participant-16 added additional comments to their response.

Question 15: "Have you observed problematic behaviors (excessive daydreaming, ambiguity and disconnection in conversations, indifference to activities, negative interactions, changes/cancellations in routines, harming oneself and others, shyness, tantrums, throwing objects) in individuals with special needs during the study?"

The findings obtained from the analysis of the answers given by the participants to Question 15 are presented in Table 16.

Table 16. *Findings related to problematic behaviors.*

Theme	Code	Frequency
Problematic behaviors	No, I did not.	18
	No, because it was a popular, interesting course.	1
	They showed no behavior to indicate such concerns	1
	Problematic behaviors were observed in some individuals with special needs	2
	The teacher intervened in a timely manner	1

Participants were asked whether they observed problematic behaviors (excessive daydreaming, ambiguity and disconnection in conversations, indifference to activity, negative interactions, changes/cancellations in routines, self-harming, shyness, tantrums, throwing objects during work, etc.) among individuals with special needs during the activity. While 18 participants answered this question by stating that they did not, only two participants (Participant-5 and Participant-19) stated that they did, albeit partially. Out of 18 participants 14 of them, who stated that they did not observe problematic behaviors among individuals with special needs, briefly replied as "No, I did not." The other four participants also made some comments on the subject. For example, Participant-10 stated that there were no problems because the activities were carried out with the support of special education teachers. Participant-15, on the other hand, emphasized that individuals with special needs completed their activities very well and without any problems. In addition, Participant-18 stated that the individual felt happier and that their previous negative behaviors had disappeared. As for the opinions of two participants who stated that they observed some problematic behaviors among individuals with special needs, Participant-5 stated that they did not witness a behavior such as throwing objects, but there was occasional disconnection, and these were fixed with the intervention of the teacher.

4. DISCUSSION and CONCLUSION

Based on the findings obtained from the research, it can be concluded that visual arts activities have a positive impact on the socialization of individuals with special needs. They are enthusiastic about participating in these activities and collaborate effectively during the activity which results in a sense of belonging. Additionally, they comply with social norms, adapt well to their environment, and are proficient in communicating with peers.

The research also revealed many positive behaviors, such as mutual respect, approval, and motivation among individuals with special needs, while no negative behavior was observed. Interestingly, individuals with special needs who often exhibit behavior problems did not display these issues during visual arts activities. Moreover, it was observed that providing an appropriate environment for expressing themselves through visual arts activities made them happy, engaged, and more sociable.

Individuals who were willing to participate in visual arts activities collaborated with their peers and teachers. Similarly, Salderay (2008), Mayer (1999), and Polloway and Patton (1997) stated in their research that individuals with special needs had difficulty integrating into society due to limitations and environmental factors and that these difficulties could be overcome through visual arts education.

It was observed that individuals with special needs who had the opportunity to express themselves through art activities socialized with each other when a suitable space was created. No aggressive behaviors were observed among individuals with special needs during visual art activities. It was observed that individuals with special needs who had difficulty expressing their feelings and thoughts, and therefore, had communication difficulties, overcame this situation during the activities. Many positive behaviors such as respecting, approving and motivating each other were observed. As a result of the opportunity to express themselves comfortably during the implementation, they found the activity enjoyable. It was concluded that individuals with special needs were eager to participate in the activities, cooperated with their peers and the teacher during the activities, developed a sense of belonging because they participated in these activities willingly, and that these individuals were effective in the realization of their adaptation to social rules and the environment, and in peer communication. At the end of the activity, teachers and individuals with special needs reported that they would like to repeat similar activities in the same area for a long time. Hetland et al. (2015) have noted that individuals whose talents are revealed through visual arts activities are likely to have a positive impact on their socialization process when their sense of achievement and self-confidence increase. Salderay (2014) has also highlighted in her research that visual arts boost self-confidence and enhance social and independent living skills in individuals.

According to Kavale et al. (1988), Rooney (2004), Keirstead and Graham (2004), Salderay (2008), and Doğutaş (2017), visual art activities allow individuals with special needs to interact with other learning areas, integrate easily with other disciplines, learn through play and fun, and engage in peer communication and cooperation. These studies share similar characteristics with the present research findings. Furthermore, the current study supports the benefits of visual art activities, including learning through play and fun, peer communication and cooperation. Many skills can be acquired through visual arts education for individuals with special needs (Grytting, 2000; Adan, 2009; Bolu, 2010; Kaynak, 1995; Işık, 2014; Paksoy, 2003). Those who develop the ability to express themselves can also enhance their motivation, conflict resolution, problem-solving, and decision-making skills (Gündoğdu & Adıgüzel, 2016). As a result of their experiences with visual art activities, it can be concluded that individuals with special needs have improved critical thinking, problem-solving, communication, and ability to cope with difficulties.

Encouraging communication and socialization among individuals with special needs has been found to increase their self-sufficiency and ability to cope with stress. Visual art activities promote emotional, mental, and social integration, and individuals have shown success in areas such as task mastery, attention and interest, sense of accomplishment, and emotional balance. Participating teachers' statements corroborated individuals' happiness in these activities, and they have not encountered them experiencing stress-inducing feelings of inadequacy or failure. Visual art activities direct individuals with special needs towards creation and production, increasing their self-sufficiency and competence. These activities have also been observed to help individuals who struggle with expressing their emotions and thoughts, overcoming communication barriers, and distancing themselves from aggressive behavior. Individuals who can express themselves have expressed a desire to repeat similar activities.

By engaging in visual art activities, individuals who acquire the ability to express themselves can use their creations as a means of communication with their environment. Simple art interventions can foster positive experiences by promoting active creativity (Martin et al., 2018). Salderay (2010) points out that visual arts are employed for therapeutic purposes and possess the power to heal individuals with special needs and patients by enhancing self-awareness and providing mental relaxation. Martin et al. (2018) mention that art activities promote socialization and active creation, leading to positive experiences. According to Salderay (2010), where visual arts are used for therapeutic purposes; they have healing power over individuals and patients with special needs, providing progress in self-perception and spiritual relaxation. Martin et al. (2018) mention that activities create positive experiences through socializing and active creation. Creative art therapies and art activities can also significantly reduce stress in participants. Scott (2017) has found that being involved in an artistic activity can reduce clinical and non-clinical stress levels.

Individuals with special needs showed increased attention and interest during the visual art activity, and no instances of failure, fear, anxiety, or restlessness were observed. It can be concluded that visual art activities have a positive impact on the behavior of individuals with special needs, improving their psychomotor skills, and reducing their levels of stress and anxiety.

Individuals with special needs did not exhibit anxiety regarding acceptance by their peers during activities, emotional instability (such as constant anxiety, excessive excitement, feeling of inadequacy, laughing or crying for no reason), or problematic behaviors (such as excessive daydreaming, uncertainty or disconnection in speech, indifference to the activity, negative interactions, changes or cancellations in routines, self-harm or harming others, shyness, tantrums, or throwing objects during work). According to Erbay (1995), it is possible to slow down the behavior of individuals who exhibit problematic behaviors through art education. Several studies, including Grytting (2000), Riccio, Rollins, and Morton (2003), Keirstead and Graham (2004), Rooney (2004), and Bayraktar (2007), have reported positive outcomes in socialization, communication, anxiety reduction, and vocational education for individuals with special needs through visual arts education. These findings suggest that individuals with special needs can feel safe and happy in society by providing visual art activities that match their level of interest and ability.

4.1. Recommendations

Based on the research findings, several recommendations have been proposed for policymakers, educators, researchers, and families. These recommendations have been categorized into two groups: those for education practice and those for future research.

4.1.1. *Recommendations for education*

1. In order to support individuals with special needs, training programs should be developed for parents and teachers in various visual art activities.
2. Teachers and families should facilitate experiential activities for individuals with special needs. They should be able to provide suitable materials based on their interests and offer feedback to promote their performance.
3. Awareness-raising campaigns should be organized to highlight the significance of visual arts in the development of individuals with special needs.
4. Special education teachers should receive training to develop a deeper interest in visual arts and institutions should employ field specialists to conduct visual arts workshops.
5. The curriculum in special education practice schools should be expanded by increasing course hours for visual arts.
6. The creation of an art space outside the workshop would enable individuals with special needs to continue their creative endeavors after class hours, contributing to their socialization and self-development.

4.1.2. *Recommendations for future research*

1. Within the scope of art education, it is possible to apply the current research to individuals with special needs studying at different grade levels, and similar studies can be conducted with individuals with special needs for different branches.
2. Studies that enable individuals with special needs and individuals with normal development to socialize with each other should be carried out.
3. Visual arts education should be emphasized in the process of accepting children from families with individuals with special needs.
4. It is important to compare domestic and foreign visual arts programs and to conduct studies that will enable individuals to focus on their areas of interest.

Acknowledgments

I would like to express my endless gratitude to my beloved wife, Barış BAYER, who supported me throughout my work, was with me at every stage, and never left me alone, and to my beloved children, who were always here with their presence during this process.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Trabzon University, Social and Human Sciences Research and Publication Ethics Committee, E-81614018-000-320.

Authorship Contribution Statement

The authors contributed equally to all the stages of the study.

Orcid

Kıymet BAYER https://orcid.org/0000-0003-1817-2022
Seda LİMAN TURAN https://orcid.org/0000-0001-6920-6558

REFERENCES

Adan, A. (2009). *Görsel sanatlar eğitiminin down sendromlu çocuklar üzerindeki etkileri (Diyarbakır örneği)* [*The effects of visual arts education on children with down syndrome (Diyarbakır Example)*] [Unpublished master's thesis]. İnönü University.

Aksu, M. (2011). *Zihinsel engelli çocuklarda görsel sanatlar eğitiminin motor beceri gelişimine katkısı* [The contribution of visual arts education to motor skill development in children with intellectual disabilities]. [Unpublished master's thesis]. Uludağ University.

Baydağ, C. (2013). *Görme engelli bireylerin sosyalleşme sürecinde verilen müzik eğitiminin, müzikal motivasyon, müziksel ilgi ve müzik yaşantılarına etkisi* [The effect of music education given in the socialization process of visually impaired individuals on musical motivation, musical interest and music experiences] [Unpublished master's thesis]. Marmara University.

Bayraktar, E. (2007). *Görsel sanatlar eğitiminin otistik çocuklar üzerindeki etkileri* [The effects of visual arts education on autistic children] [Unpublished master's thesis]. Gazi University.

Bolu, H. (2010). *Zihinsel engelli çocukların gelişiminde sanat eğitimi derslerinin katkısı* [*Contribution of art education courses in the development of mentally retarded children*] [Unpublished doctoral dissertation]. Marmara University.

Cüceloğlu, D. (1994). *İnsan ve davranışı: Psikolojinin temel kavramları* [Humans and their behavior: Fundamental concepts of psychology]. Remzi Books.

Doğutaş, A. (2017). *Özel eğitim kurumlarındaki görsel sanat eğitim uygulamalarının öğretmen görüşleriyle değerlendirilmesi* [*Evaluation of visual art education practices in private education institutions based on the opinions of teachers*] [Unpublished master's thesis]. Yakın Doğu University.

Eisner, E.W. (2002). What can education learn from the arts about the practice of education? *Journal of curriculum and supervision*, *18*(1), 4-16.

Erbay, M. (1995). *Yükseköğretim düzeyinde sanat eğitimi programlarının uluslararası bağlamda incelenmesi* [Examination of arts education programs at the higher education level in an international context]. [Unpublished doctoral thesis]. Marmara University.

Eşsizoğlu, A., Işıklı, B., Güleç, G., Aksaray, G., Yenilmez, Ç., & Kırel, A.Ç. (2013). *Çatışma ve stres yönetimi-II* [Conflict and stress management-II]. Anadolu University Publication. https://ets.anadolu.edu.tr/storage/nfs/CMH202U/ebook/CMH202U-13V4S1-8-0-1-SV1-ebook.pdf

Grytting, C. (2000). The benefits of art education. Forum: Thoughts to Share. *Arts & Activities*, *127*(3), 66.

Gündoğdu, R., & Adıgüzel, Ö. (2016). Stres ve yaratıcı drama: Üniversite öğrencileri ile yapılan bir çalışma [Stress and creative drama: A study with university students]. *Creative Drama Magazine*, *11*(1), 45-70.

Hetland, L., Winner, E., Veenema, S., & Sheridan, K.M. (2015). *Studio thinking 2: The real benefits of visual arts education*. Teachers College Press.

Işık, H. (2014). *Zihinsel engelli çocukların el becerilerinin gelişmesine resim-iş dersinin katkısı ile ilgili özel eğitim öğretmenlerinin görüşleri* [*Opinions of special education teachers about the contribution of art lessons to the development of hand skills of mentally handicapped children] [Unpublished doctoral dissertation*]. Anadolu University.

Keirstead, C., & Graham, W. (2004). *VSA arts research study: Using the arts to help special education students meet their learning goals*. RMC Research Corporation.

Martin, L., Oepen, R., Bauer, K., Nottensteiner, A., Mergheim, K., Gruber, H., & Koch, S.C. (2018). Creative arts interventions for stress management and prevention-a systematic review. *Behavioral Sciences*, *8*(2). https://www.mdpi.com/2076-328X/8/2/28/pdf

Mayer, H. (1999). *Kompetenzprofil; Staatlich Anerkannte Arbeitserzieherin Staatlich Anerkannter Arbeitserzieher*. Schemmerhofen.

Meros, D. (1990). *A developmental/behavioral to art therapy for persons with mental retardation* [Doctoral dissertation]. Ursuline College.

Paksoy, S. (2003). *8-12 yaş eğitilebilir zihinsel engelli çocukların resim-iş eğitimi* [*Art-work education of educable mentally retarded children aged 8-12*] [Unpublished doctoral dissertation]. Marmara University.

Polloway, E.A., Patton, J.R., Smith, T.E., & Buck, G.H. (1997). Mental retardation and learning disabilities: Conceptual and applied issues. *Journal of Learning Disabilities*, *30*(3), 297-308. https://doi.org/10.1177/002221949703000305

Rooney, R. (2004). *Arts-based teaching and learning: A review of the literature*. VSA Arts. http://www.kennedy-center.org/education/vsa/resources/VSAarts_Lit_Rev5-28pdf

Salderay, B. (2008). *Türkiye'deki zihin engelliler iş okullarında görsel sanatlar dersinin öğrencilerin beceri, davranış ve meslek edinimindeki katkısına yönelik öğretmen görüşleri* [Teachers' views on the contribution of visual arts course to students' skill, behavior and vocational acquisition in workshop schools for the mentally handicapped in Turkey] [Unpublished doctoral dissertation]. Gazi University.

Salderay, B. (2010). Engelli bireylerin yapmış olduğu görsel sanatlar çalışmalarının engelli birey aileleri ve engelli bireylerle çalışan eğitimciler tarafından değerlendirilmesi [Evaluation of visual arts works by disabled individuals by their families and educators working with disabled individuals]. *Journal of Erzincan Faculty of Education, 12*(1), 219-229.

Salderay, B. (2014). Özel eğitim sürecinde görsel sanatlar uygulamalarının önemine ilişkin aile düşünce yapılarının değişimi [Change of family mentality regarding the importance of visual arts practices in the special education process]. *Manas Journal of Social Studies, 3*(3), 87-101.

Scott, B.A. (2017). *Art as a stress reduction tool* [Unpublished doctoral thesis]. Marietta College.

Şahan, H. (2007). *Üniversite öğrencilerinin sosyalleşme sürecinde spor aktivitelerinin rolü* [The role of sports activities in the socialization process of university students] [Unpublished doctoral thesis]. Selçuk University.

Taş, R. (2019). *Ortaokul öğrencilerinde sosyal etkinliklere katılma durumu ile stres düzeyleri arasındaki ilişkinin incelenmesi* [Examining the relationship between participation in social activities and stress levels in secondary school students] [Unpublished master's thesis]. Biruni University.

Yıldırım, A., & Şimşek, H. (1999). *Sosyal bilimlerde nitel araştırma yöntemleri* (11. bs.) [Qualitative research methods in social sciences (11th ed.)]. Seçkin Publications.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 709–722

https://doi.org/10.21449/ijate.1271693

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Comparison of cronbach's alpha and McDonald's omega for ordinal data: Are they different?

Fatih Orcan[1,*]

[1]Kahramanmaraş Sütçü İmam University, Faculty of Education, Department of Educational Sciences, Türkiye

ARTICLE HISTORY

Received: Mar. 27, 2023
Revised: Oct. 03, 2023
Accepted: Oct. 16, 2023

Keywords:
Cronbach's Alpha,
McDonald's Omega,
Reliability,
MonteCarloSEM,
Simulation.

Abstract: Among all, Cronbach's Alpha and McDonald's Omega are commonly used for reliability estimations. The alpha uses inter-item correlations while omega is based on a factor analysis result. This study uses simulated ordinal data sets to test whether the alpha and omega produce different estimates. Their performances were compared according to the sample size, number of items, and deviance from tau equivalence. Based on the result, the alpha and omega had similar results, except for the small sample size, the smaller number of items, and the low factor loading values. When there were 5 or more items in the scale and factor analysis which the omega was calculated from showed fit to the data set, using omega over alpha could be preferred. Also, as the number of items exceeds 5, the alpha and omega differences disappear. Since calculating the alpha is easier compared to the omega (omega requires fitting a factor model first) using alpha over omega can also be suggested. However, when the number of items and the correlations among the items were small, omega performed worse than alpha. Therefore, alpha should be used for the reliability estimations.

1. INTRODUCTION

One of the most critical steps for a scientific study is data collection. This is also critical for reliability. Thus, the data collection process is better planned in order to get information as reliably as possible. Reliability is a property of the data collected, not the scale instrument itself (Streiner, 2003). Therefore, the data collected from the same instrument could be reliable for one example and not for another. One way of getting reliability is doing the test-retest procedure. According to the test-retest method, the same test is administered to the same group of examinees twice at a time interval. The correlation between the two-test administration is called test-retest reliability. Another method, which is commonly used (Streiner, 2003; Vaske, et al., 2017), is Cronbach's alpha (α). The alpha is also known as internal consistency reliability since it uses inter-item correlations to calculate reliability. The higher the correlations more reliable the data. The alpha value is also a property of the data. When the data changes, the alpha will also change. For reliable data collection, the change in the alpha is expected to be small compared to previous ones. However, as Henson et al. (2001) pointed the alpha values for the same instrument (a.k.a., Internal failure scale) ranged between .51 and .82 (as cited in Streiner, 2003).

*CONTACT: Fatih Orçan ✉ fatihorcan@ksu.edu.tr Kahramanmaraş Sütçü İmam University, Faculty of Education, Department of Educational Sciences, Kahramanmaraş, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

Cronbach's alpha comes with a few assumptions (Kalkbrenner, 2023). First of all, the items under the scale should be unidimensional (Edwards et al., 2021) which means all of the items are related only to one latent construct. Second, the items should be normally distributed and continuous (Edwards et al., 2021). Also, the error terms of the items should not be correlated. That is, there should not be any other common variance among the items except the latent factor. Last but not least, the items in the scale should equally contribute to the latent factor, the essentially tau equivalence assumptions. In case the assumption is not satisfied the alpha values underestimate the true reliability (Edwards et al., 2021; Kalkbrenner, 2023).

There are more than 30 reliability calculation methods (Edwards et al., 2021). Besides Cronbach's alpha, McDonald's omega (ω) is one of them. The alpha and omega values showed the most accurate estimate of reliability (Edwards et al., 2021). The omega reliability does not have such assumptions as the alpha. Therefore, when Cronbach's alpha does not hold its assumptions, it is recommended not to use alpha but to use omega instead (Goodboy & Martin, 2020). Specifically, when the tau equivalence assumption does not hold, use of ω is suggested (Viladrich et al., 2017). It is because, McDonald's omega is robust to the violations of the assumptions (Goodboy & Martin, 2020; Kalkbrenner, 2023). In fact, the alpha and omega give equally good results if the assumptions are met (Edwards et al., 2021; Goodboy & Martin, 2020; Viladrich et al., 2017).

The omega estimations are based on confirmatory factor analysis (CFA). A CFA model fits the data first and then the omega is calculated based on the factor loadings and the error variances as given in the formula:

$$\omega = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \theta_i}$$

where the λ_i represents the factor loadings for item i, and θ_i represents the error variance of the item.

Even though omega does not have such assumptions as the alpha, since omega is calculated after a CFA, anything affecting model-data fit for the CFA model also affects the omega value. For example, the sample size is a critical issue for a factor analysis. As the sample size gets lower, the model-data fits become problematic for factor analyses or even the model may not converge to a solution (Gagne & Hancock, 2006). Also, under small sample sizes and unequal factor loadings omega estimation becomes biased (Edwards et al., 2021). Also, the number of items in the structure affects the omega reliability. Increasing the number of items stabilizes the omega estimates even for small sample sizes (Edwards et al., 2021; Ercan et al., 2017). In short, even though ω is a good alternative to α, the alpha produced more accurate estimates under a small sample size and number of items (Edwards et al., 2021).

The alpha and omega estimates also differ when the factor loadings have different values (e.g., non-tau equivalence) under a factor analysis (Edwards et al., 2021). When there is a discrepancy among the factor loading and as the size of the discrepancy increases, the alpha and omega produce different results. However, the difference between alpha and omega has "no practical consequences" when the average factor loadings are .7 or higher and the difference among the loading values is .2 in absolute values (Raykow & Marcoulides; 2015; Viladrich et al., 2017).

1.1. Aim of the Study

Edwards et al. (2021) compared six different reliability estimations in their work and based on the results, alpha and omega produced the most accurate estimate of the true reliability. Even though the alpha and omega were shown to be better, it was also shown that each estimation has its flaws. For example, "omega was affected greater by the number of items when reliability was low" (p. 1111). However, the work of Edwards et al. (2021) was based on continuous scaled data. In applications, ordinal scaled data are often used. Seeing the performance of alpha and omega with ordinal scaled data is informative and, therefore, this paper aims to estimate alpha and omega reliabilities based on a five-point Likert-type ordinal scale as similar conditions with Edwards et al. (2021).

To compare the alpha and omega estimates data were simulated under different conditions: sample size, balanced and un-balanced factor loadings, and number of items.

2. METHOD

2.1. Data Generation Procedure

Data were simulated via *MonteCarloSEM* package (Orçan, 2021) in R-Cran (R Core Team, 2014). The *MonteCarloSEM* package simulates and analyzes data based on a given CFA model. It can produce normal/skewed or continuous/ordinal scale data sets for given threshold values. Various simulation conditions were considered in this study. First of all, the sample size varied from 50 to 1000 to represent the sample size from small to large (e.g., 50, 100, 300, 500, 800, 1000). Second, the number of items under the scale differed. The minimum number under the scale was 3 and increased to 5, 8, 10, and 20. Therefore, five different number of items were considered for this study. The minimum number of items was chosen to be 3 because it is a prerequisite for a single-factor CFA model. Also, the maximum was 20 since the correlation between the number of items and reliability was shown to be reduced after 19 items (Vaske et al., 2017). Finally, average factor loading values were also changed for this study. Under this condition, five different scenarios were tested.

- Tau: All factor loadings were equal across the factor and the average loadings were set to be .3, .4, .5, .6, .7, .8, and .9.
- Mixed-1: The loadings values were differed by .2 at maximum and average loadings were set to be .3, .4, .5, .6, .7, .8, and .9. Therefore, for the average of .3, the factor loadings were .2, .3 and .4 under three items and .2, .25, .3, .35, and .4 repeated each for four times under 20 items.
- Mixed-2: The loadings values were differed by .4 at maximum and the average loadings were set to be .4, .5, .6, and .7. For the average of .4, under three items, the loadings were .2, .4 and .6 and under twenty items the loadings were .2, .3, .4, .5, and .6 each repeated for four times.
- Mixed-3: The loadings values were differed by .5 at maximum and the average loadings were set to be .5, .6, and .7. For the average of .5, under three items, the loadings were .25, .5, and .75 and under twenty items the loadings were .25, .35, .5, .65, and .75 each repeated for four times.
- Mixed-4: The loadings values were differed by .6 at maximum and the average loadings were set to be .5 and .6. For the average of .5, under three items, the loadings were .2, .5, and .8 and under twenty items the loadings were .2, .3, .5, .7, and .8 each repeated for four times.

For all the scenarios, the loadings were increased by .1 each time to increase the average factor loadings, respectively. For example, under the mixed-4 condition, to get average loadings to be .6 for three items the loadings were set as .3, .6, and .9.

2.2. Data Analysis

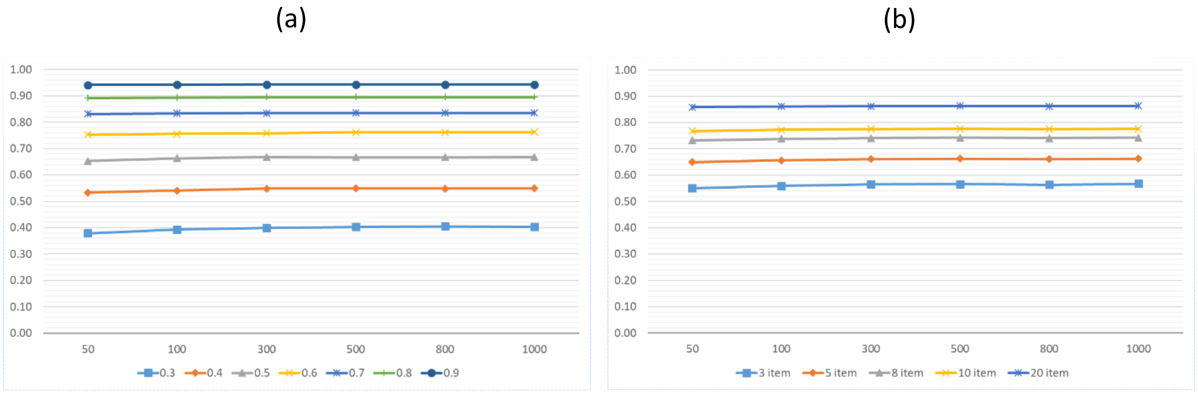
Each condition was repeated 1000 times by using the *sim.categoric()* function in the *MonteCarloSEM* package. Ordinal data were created by using -1.645, -.643, .643, and 1.645 as the threshold values to create a 5-point Likert scale. The Cronbach's alpha values were calculated by using *CronbachAlpha()* function in the *DescTools* package (Signorell, 2023). To estimate omega values one-factor CFA models were fitted to the simulated data sets by the *cfa()* function in the *lavaan* package (Rosseel, 2012), using the maximum likelihood estimation method. Besides the omega, model-data convergence rates and fit indices such as the p-value of the chi-square test, the comparative fit indices (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) were saved for the further evaluations. Hu and Bentler's (1999) criteria were used to evaluate the fit. Moreover, the relative biases where the absolute difference between the true and estimated values were divided by the true value were calculated:

$$Relative\ Bias = \frac{Abs(True\ Value - Estimated\ Value)}{True\ Value}$$

3. RESULTS

Based on the results of the simulation, the sample size was not an important factor for the Cronbach alpha estimation, as expected. The average estimates were almost identical across different sample sizes. As the sample size was increased from 50 to 1000, the alpha estimates changed only by .02 for the low factor loading values (.3, .4, and .5). When the factor loading values were increased, the gap disappeared and the estimates became identical. Similarly, when the number of items was 3, the average alpha estimates differed by .02, and as the number of items got larger, the gab became .01. Figure 1 showed the change of average alpha estimates by sample size for different values of the factor loadings and the number of items. As expected, the estimates increased by the increase in the factor loadings and the number of items. However, the estimated values had horizontal lines across different sample sizes.

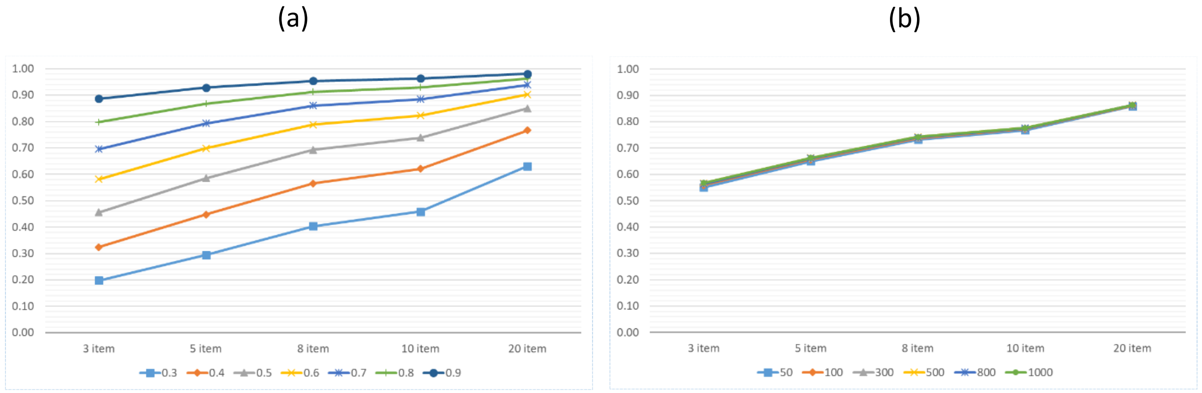
Figure 1. *Alpha estimated by sample size for factor loadings (a) and number of items (b).*



Different from the sample size, the number of items affected the alpha estimates according to the simulation results. Figure 2 shows the change of average alpha estimates by the number of items for different values of factor loadings and sample sizes. As the total number of items was increased, the estimates also increased. This was the result expected, since the effect of the number of items on the alpha is well-known in the literature (Streiner, 2003; Vaske et al., 2017). In addition, the effects of the number of items on the alpha estimates were larger, when the factor loadings were smaller (See Figure 2/a). For example, when the factor loading was .3 and the number of items was increased from 3 to 20, the estimate jumped to .63 from .20. However, when the factor loading was .6, the change in the estimates was smaller, from .58 to .90. The gap even got smaller when the factor loading was .9, from .89 to .98.

Based on the results of the simulation, under the model where the factor loadings were equal (Tau model), the α and ω values differed only for small sample sizes (50 and 100), small factor loadings (.3 and .4), and less number of items (3 and 5). The difference between the values ranged between .06 and .18. For all other conditions under the tau models difference was smaller than .04 and as the sample size, number of items, and factor loadings were increased, the gaps disappeared. Figure 3 shows the alpha (α) and omega (ω) estimates as well as the relative biases and the true values under the tau models for sample sizes of 50, 300, and 1000 (only three sample sizes were given due to space limitations).

Figure 2. *Alpha estimated by number of items for factor loadings (a) and sample sizes (b).*



For example, when there were 3 items, the loadings were all set to be .3, and the sample size of 50, the alpha and omega estimates were .18 and .35, respectively where the true value was .23. However, when holding all other values constant but the sample size was increased to 100, the values become .19 and .33. On the other hand, when the number of items becomes 5 instead of 3, the values become .27 and .39, indicating a .12 difference. Similarly, when factor loadings were increased to .4, the values became .31 and .45. Besides, the omega estimates were always larger than the alpha, except for sample sizes of 50, 20 items, and factor loadings of .3. Under this condition, the alpha and omega values were .62 and .61, respectively.

Figure 3 also showed relative biases for α and w. Based on Bandalos's (2002) recommendation, which pointed that relative bias should be smaller than .10, the relative biases pointed to problematic values under small sample size (50 and 100), small factor loadings (.3 and .4), or a smaller number of items (3 and 5). That is, as the sample size, number of items, and/or average factor loadings were increased, the relative biases decreased and got under the .10 critical value. Interestingly, under these conditions, alpha and omega estimates pointed to almost identical relative biases. For example, the relative biases showed similar values even when the sample size was as small as 50, the number of items 8, and factor loadings .4. Increasing values of any of these simulation conditions diminished the relative biases and the gap between the alpha and omega estimates.

Almost identical results with the Tau model were obtained for the mixed factor loadings model where the factor loading differs only by .2 (Mixed-1 model). Therefore, when tau equivalence was not granted and the difference between the loadings was up to .2, using omega reliability instead of the alpha did not change the results, except for small sample size (50 and 100), small factor loadings (.3 and .4) and less number of items (3 and 5) as it was the case for the tau models. That is to say, even for a small sample size the alpha and omega were almost identical (the difference was at the third decimal) as long as the factor loading differs by .2, the average factor loadings were above .5, and the number of items was more than 5. Figure 4 shows the alpha (α) and omega (ω) estimates for Mixed-1 models for the sample sizes of 50, 300, and 1000. Based on Mixed-1 model results, when the sample size was increased, the gap between the alpha and omega estimated disappeared even for small factor loadings and/or fewer items. For example, when the sample size was 300, the number of items was 3, and the average factor loading was .5, the difference between alpha and omega was only .02. Keeping everything constant but increasing the number of items to 5, the difference disappeared (.006).

Figure 3. *Alpha and omega estimates and relative biases for tau models under different sample sizes.*

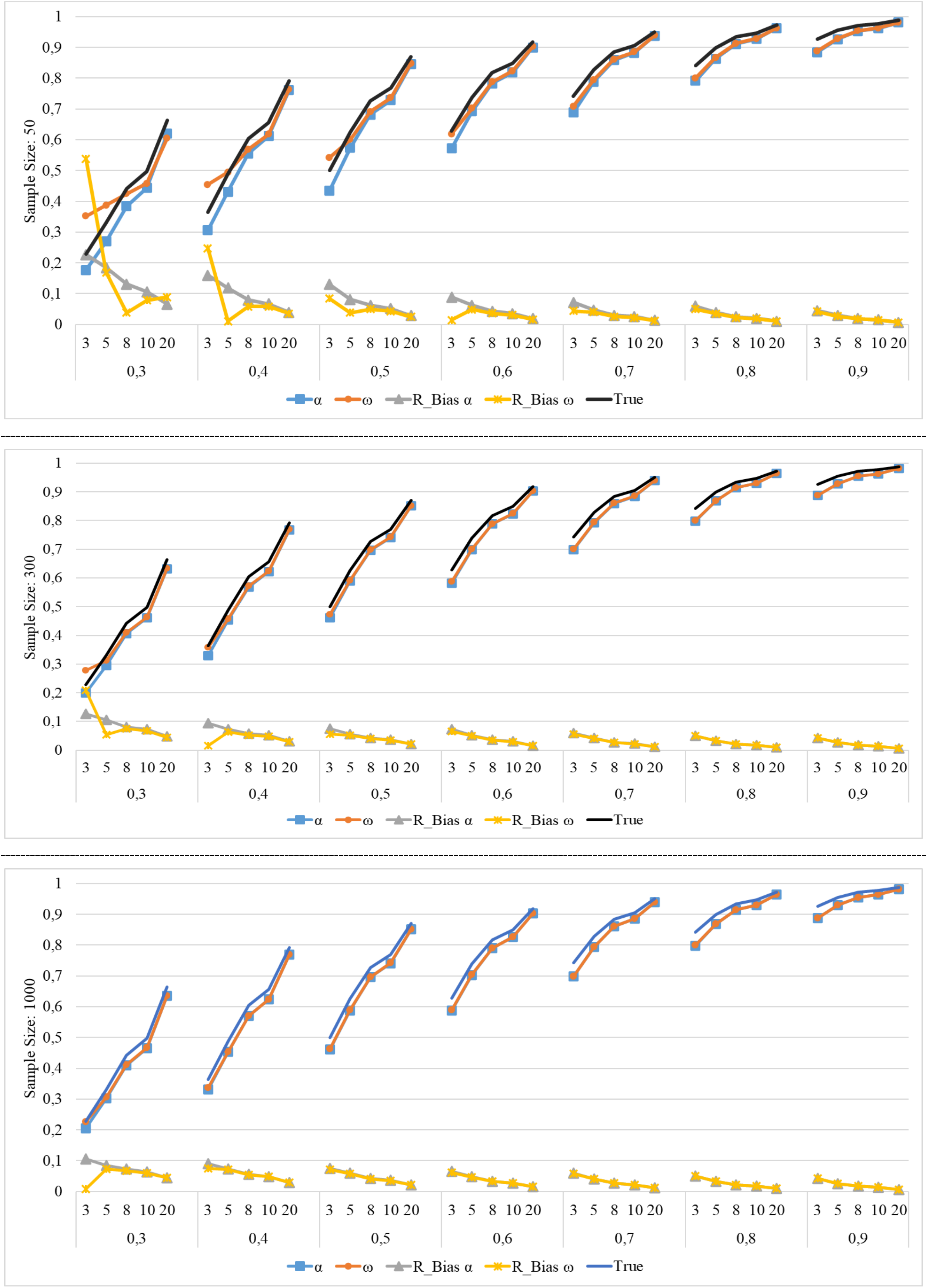


Figure 4. *Alpha and omega estimates and relative biases for mixed-1 models under different sample sizes.*

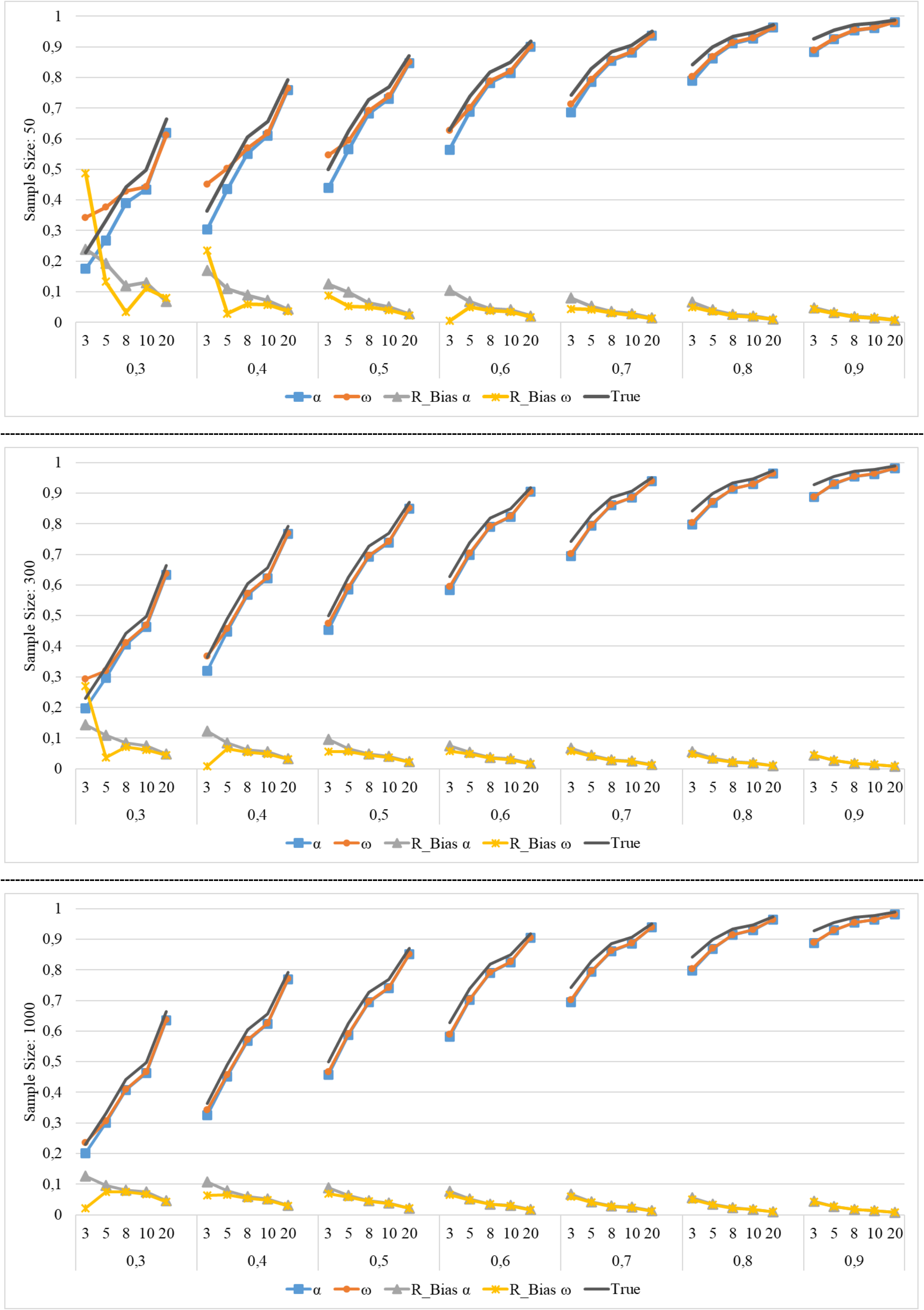


Figure 5. *Alpha and omega estimates and relative biases for mixed-2 models under different sample sizes.*

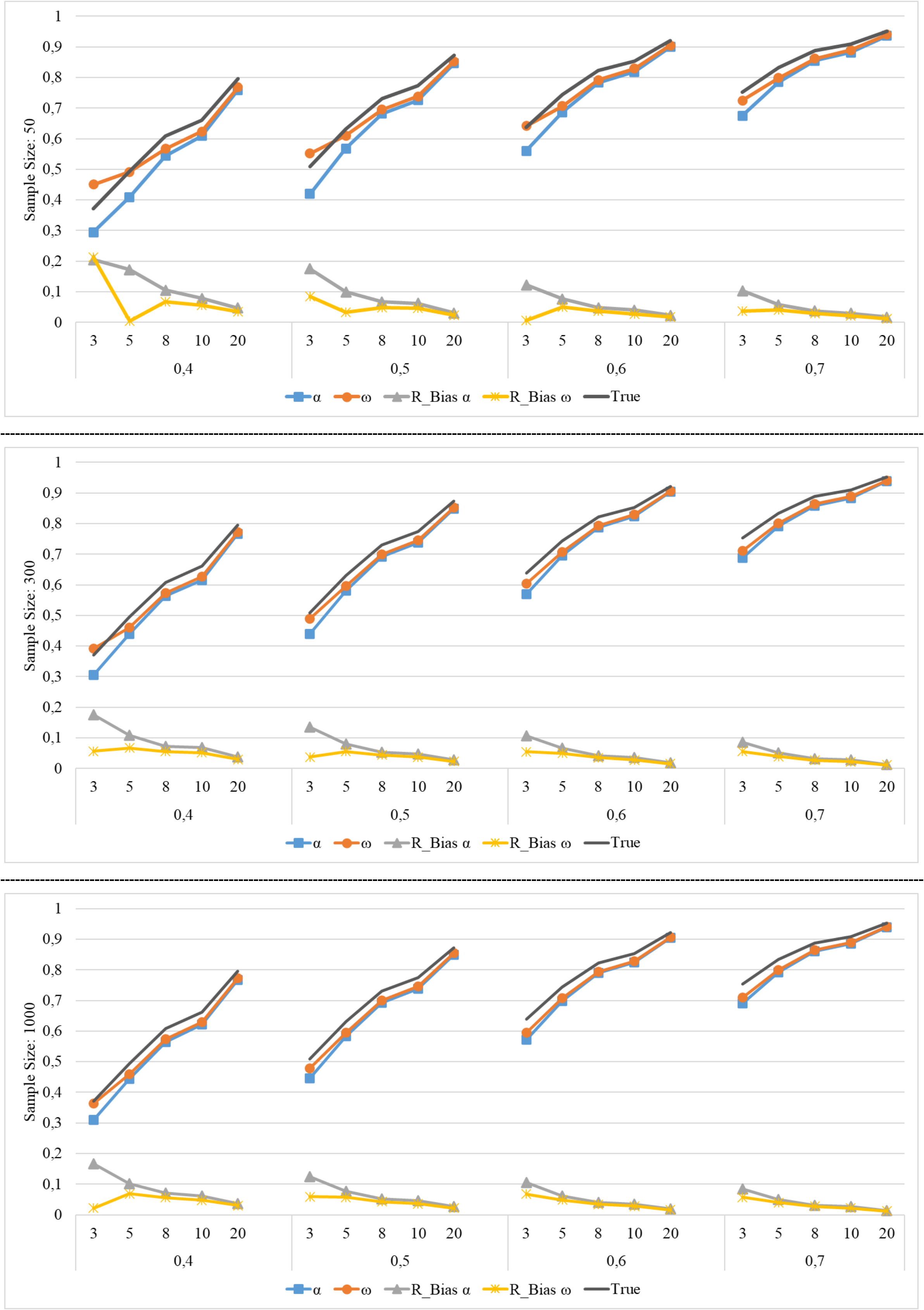
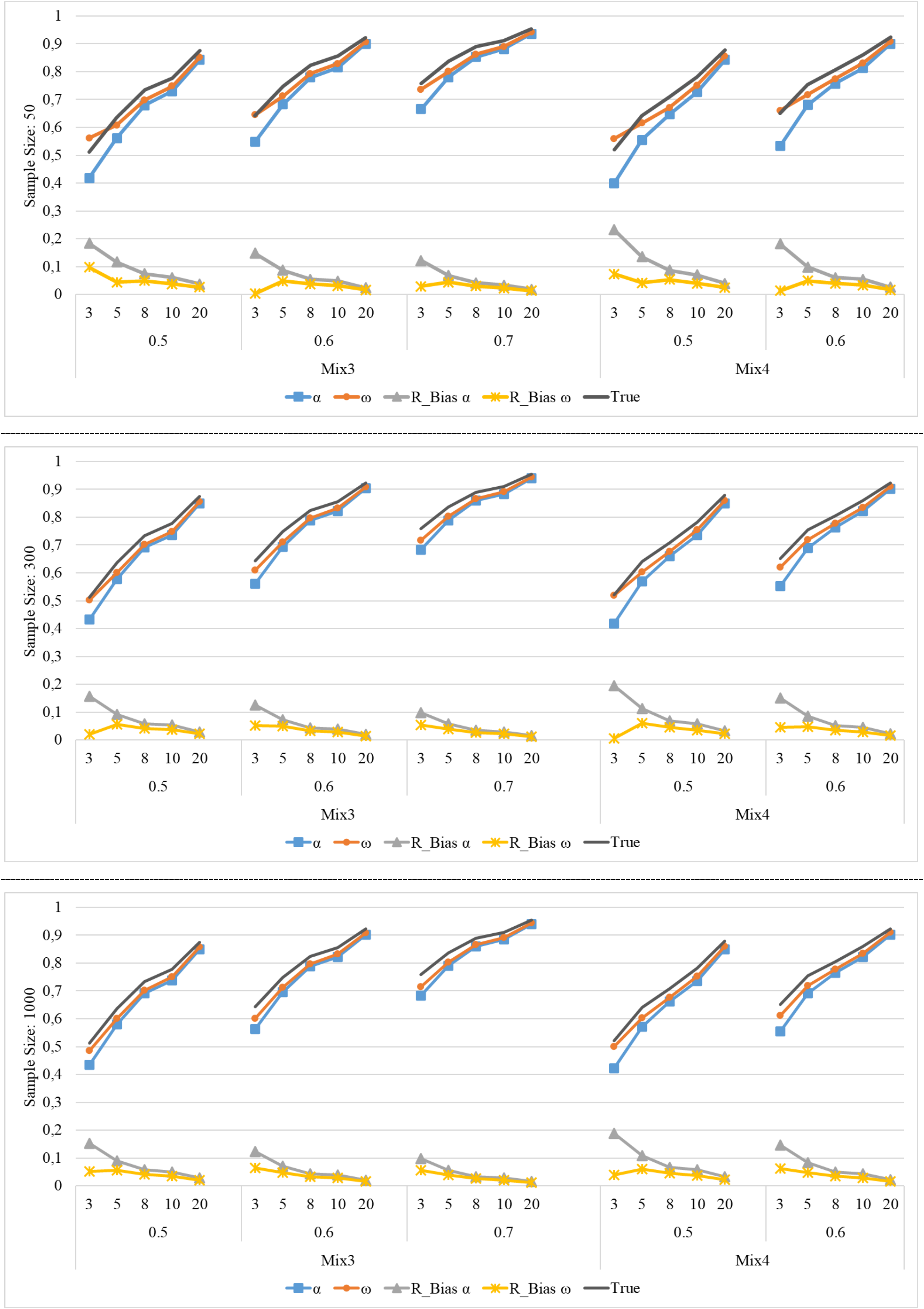


Figure 6. *Alpha and omega estimates and relative biases for mixed-3 and mixed-4 models under different sample sizes.*



Similar results were produced among the other mixed factor loadings (a.k.a., non-tau) models where the factor loading differs by .4 (Mixed-2 model), the factor loading differs by .5 (Mixed-3 model) and the factor loading differs by .6 (Mixed-4 model). The results of the models are reported in Figure 5 and Figure 6. When the number of items was 3 and average factor loadings were smaller than .6, all these conditions produced considerable gaps between the alpha and omega estimates, even for a sample size of 1000. That is, when there were 3 items at a scale and the average of factor loadings was smaller than .6, the tau equivalence became important under the Mixed-2, 3, and 4 models. Even though the gap got smaller with the sample size for some conditions, the change was smaller than .02.

Therefore, based on the results of the study, using alpha or omega to estimate reliability becomes not an important issue when the sample size and the number of items are larger than 300 and 3, respectively. The average of factor loadings and the difference between the factor loadings at a scale become important only when the number of items is 3.

The omega estimates were based on factor analysis results. Therefore, a CFA model should be tested, and the results should show a good model-data fit first. According to the simulation results, when the sample size and factor loadings were small, under 3 and 5 items models, the CFA models showed a higher percent of convergence problems. In other words, when there were more than 5 items, the average factor loadings were larger than .5, and the sample size was larger than 100, the model convergence was not a problem. Figure 7 shows the percentage of non-convergences under different sample sizes (Only the results of the sample sizes of 50 and 100 were given since as the sample size increased, they had no convergences). For example, when the sample size was 50, the number of items was 3, and the average factor loading was .3 under Tau and Mixed-1 models, about 28% of the data did not converge to a solution. The non-convergence rates decreased as the sample size, the number of items, and the average factor loadings were increased. Also, when the sample size and factor loadings were small, under 3 and 5 items models, the supplementary fit indices (CFI, RMSEA, and SRMR) indicated model-data fit issues too, in case the model converged to a result.

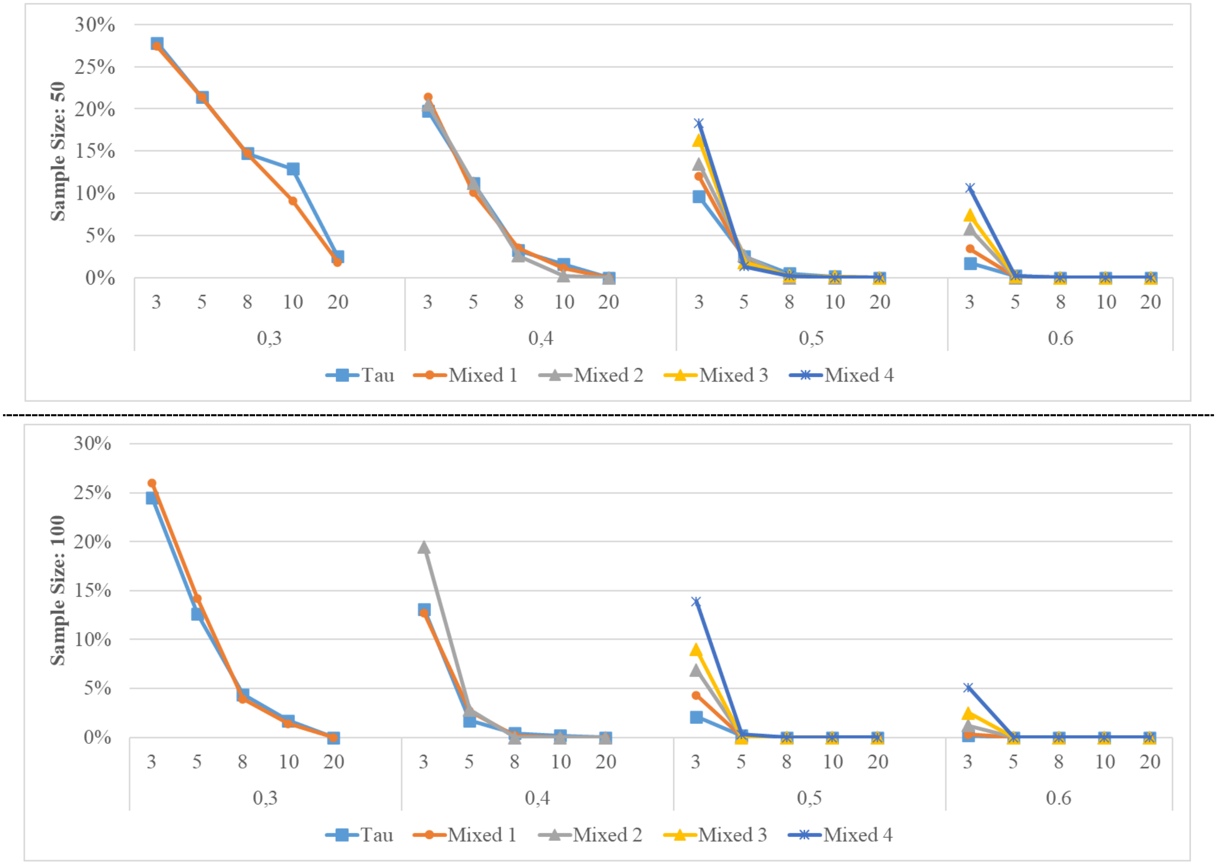
4. DISCUSSION and CONCLUSION

Among others, Cronbach’s Alpha (α) and McDonalds’ Omega (ω) were used commonly for the reliability estimates. Also, it was shown that the alpha and omega produced the most accurate and similar estimates of reliability (Edwards et al., 2021). Therefore, only these two reliability estimates were considered in this study. Based on the results of the current study, α and ω had similar results, except for the small sample size, a smaller number of items, and low factor loading values. Since the omega estimates were based on CFA results and factor analysis requires a larger sample size to converge a solution, under small sample sizes, the ratio of convergence was low. Therefore, the gap between the estimates of α and ω might be due to the convergence problem. Also, as the convergence rate increased with sample size, number of items, and factor loading, the gap between the estimates of α and ω got smaller and disappeared eventually.

Related to the convergence problem, even when the model converged to a solution, supplementary fit indices (CFI, RMSEA, and SRMR) sometimes indicated problems regarding the model-data fit. Similar to the convergence problem, the fit indices showed problems only when the sample size was small, the number of items and the factor loading values were low. Therefore, in case the data fit creates a concern, the estimated omega values might be problematic and “should not be used” to estimate reliability (McDonald, 2011, p. 89).

Relative biases were also calculated for the estimates (See Figures 3 to 6). When the number of items was larger than 5, relative biases were almost identical under all models (Tau, Mixed-1, 2, 3, and 4), regardless of the average factor loadings or the sample sizes.

Figure 7. *Percent of non-convergences under different sample sizes.*



When the number of items was 3, the sample size was small and the average factor loadings were .4 and smaller, omega showed greater bias than alpha. However, when the number of items was increased to 5, the bias of omega was smaller than that of alpha. Therefore, similar to the convergence issue, it can be concluded that when sample size, number of items, and average factor loadings were small, the omega estimates became worse than the alpha estimates. However, under all other conditions, ω produced less or equal relative biases than α. Therefore, when there are 5 or more items in the scale and the CFA model fits to the data set, using ω over α could be preferred since it produced less or equal bias compared to the α. It was also worth pointing that as the number of items becomes larger than 5 the difference between the estimates becomes smaller. Therefore, under these conditions, using alpha or omega to estimate reliability does not affect the results. From this point of view, since calculating the alpha is easier compared to omega because omega requires fitting a CFA model first, using alpha over omega can also be suggested.

Under mixed-2, mixed-3, and mixed-4 models, since average factor loadings were all larger than .4 due to the design factors, the omega only outperformed the alpha estimates when the number of items was 5 or smaller. Especially when the sample size and number of items were small, the alpha produced relative biases larger than the critical value (10%). If there were more than 5 items, the alpha and omega estimates were almost indistinguishable. To conclude, when there were more than 5 items, alpha and omega produced similar results, regardless of sample size, average factor loadings, or the tau equivalence assumptions. Therefore, based on the results, using alpha to estimate reliability is not wrong. However, when the CFA model, which the omega is calculated from, fits the data well, using omega to estimate reliability is also reasonable.

In conclusion:

- Even with tau models, alpha produced biased results under the small number of item number, sample size, and average factor loadings. However, under the same conditions, omega produced more bias. Similarly, Edwars et al. (2021) reported that omega was affected more by the sample size and average factor loadings. Even though it was not reported in the study, all the biases produced by alpha were positive, indicating alpha underestimates the true value while omega produced negative biases for the small sample sizes, fewer items, and low factor loadings.
- Based on the results of the study, it is true that the further away from the tau, the more biased the alpha is compared to the omega. Similar results were also reported by Edwards et al. (2021). However, as values of the design factors increase, regardless of the sample size, alpha and omega yield similar results. Thus, as long as the values of the design factors are not small, there is no harm in using the alpha for reliability.
- The results again confirmed that alpha under-estimate the reliability. However, it only happens as the model deviates from the tau equivalence and has a smaller number of items and average factor loadings. Under other conditions, the differences between alpha and omega estimates are less than 3%.
- To estimate omega, a CFA model is required to run first. If the CFA model does not fit to the data, the omega obtained from it may also be biased. However, this is not the case for the alpha estimates. Alpha gives a result even when the CFA produces problematic results for the omega estimates (e.g., smaller sample sizes, average factor loadings, and number of items). In fact, the bias of the alpha is even lower compared to omega under these conditions. Similarly, Edwards et al. (2021) have shown that alpha was superior to omega under these conditions.
- In the case of small sample size, number of items, and low factor loadings, omega estimates showed a much larger bias compared to the bias of alpha. As the values of the design factors were increased the bias of omega and alpha became closer. The larger omega bias was most probably due to the convergence problems of the CFA model. Since the models do not converge, the results were not stable and showed deviated estimates.
- When the number of items is 3 and the correlations among the items are low, even if the sample size is 1000 and the tau equivalence holds, the alpha estimates show biased results (over 10%). However, it should be remembered that there might be convergence and fit problems for the omega under similar conditions. Therefore, if there are low correlations among the items, increasing the number of items is important to get a more accurate reliability prediction.
- Even though, it seems that both α and ω produce biased results when the sample size is small, keeping the sample size constant and increasing the average factor loadings and the number of items, the bias disappears. Despite that Edwards et al. (2021) reported the effect of sample size on the omega estimates, this study showed that it is not directly related to the omega itself. It is most probably related to the requirements of CFA models. Since a CFA model needs a larger sample size for stable estimations, in case the sample size was low, the omega might not be estimated correctly. Therefore, it cannot be said that the sample size had a direct effect on the biases. After all, Yurdagül (2008) showed that α produces unbiased estimates even when the sample size is as low as 30. In a similar situation, the high bias of ω can be associated with the model data fits.

The results were limited to the simulation conditions. Under this study, one-factor CFA models were considered. Similar comparisons of reliability estimations can also be made for multi-dimensional (a.k.a., two or more factor CFA models) structures.

Declaration of Conflicting Interests and Ethics

The author 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

Fatih Orçan https://orcid.org/0000-0003-1727-0456

REFERENCES

Bandalos, D.L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*(1), 78–102. https://doi.org/10.1207/S15328007SEM0901_5

Bernardi, R.A. (1994). Validating research results when cronbach's alpha is below .70: A methodological procedure. *Educational and Psychological Measurement, 54*(3), 766–775. https://doi.org/10.1177/0013164494054003023

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Edwards, A.A., Joyner, K.J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement*, *81*(6), 1089-1117. https://doi.org/10.1177/0013164421994184

Ercan, I., Yazici, B., Sigirli, D., Ediz, B., & Kan, I. (2007). Examining cronbach alpha, theta, omega reliability coefficients according to sample size. *Journal of Modern Applied Statistical Methods*, *6*(1), 291-303. https://doi.org/10.22237/jmasm/1177993560

Gagne, P., & Hancock, G.R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65-83. https://doi.org/10.1207/s15327906mbr4101_5

Goodboy, A.K., & Martin, M.M. (2020). Omega over alpha for reliability estimation of unidimensional communication measures. *Annals of the International Communication Association*, *44*(4), 422-439. https://doi.org/10.1080/23808985.2020.1846135

Henson, R.K., Kogan, L.R., & Vacha-Haase, T. (2001). A reliability generalization study of the teacher efficacy scale and related instruments. *Educational and Psychological Measurement*, *61*(3), 404-420. https://doi.org/10.1177/00131640121971284

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. https://doi.org/10.1080/10705519909540118

Kalkbrenner, M.T. (2023). Alpha, omega, and h internal consistency reliability estimates: Reviewing these options and when to use them. *Counseling Outcome Research and Evaluation*, *14*(1), 77-88. https://doi.org/10.1080/21501378.2021.1940118

McDonald, R.P. (2011). *Test theory: A unified treatment*. Routlege.

Orçan, F. (2021). MonteCarloSEM: An R package to simulate data for SEM. *International Journal of Assessment Tools in Education*, *8*(3), 704-713. https://dergipark.org.tr/en/pub/ijate/issue/62753/804203

R Core Team. (2014). *R: A language and environment for statistical computing* [Computer software manual]. http://www.R-project.org/

Raykov, T., & Marcoulides, G.A. (2015). A direct latent variable modeling based method for point and interval estimation of coefficient alpha. *Educational and Psychological Measurement*, *75*(1), 146–156. https://doi.org/10.1177/0013164414526039

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Signorell, A. (2023). DescTools: Tools for descriptive statistics. R package version 0.99.48. https://CRAN.R-project.org/package=DescTools

Streiner, D.L. (2003). Starting at the Beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. https://doi.org/10.1207/S15327752JPA8001_18

Vaske, J.J., Beaman, J., & Sponarski, C.C. (2017). Rethinking internal consistency in cronbach's alpha. *Leisure Sciences, 39*(2), 163-173. https://doi.org/10.1080/01490400.2015.1127189

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología / Annals of Psychology*, *33*(3), 755–782. https://doi.org/10.6018/analesps.33.3.268401

Yurdagül, H., (2008). Minimum sample size for cronbach's coefficient alpha: A Monte-Carlo Study. *H.U. Journal of Education*, *35*, 397-405. http://www.efdergi.hacettepe.edu.tr/shw_artcl-571.html

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 723–740

https://doi.org/10.21449/ijate.1311600

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Development of self-regulation scale for middle school students: Validity and reliability study

Ismail Sarikaya[1,*], Mesut Ozturk[2], Mustafa Ozgol[3]

[1]Bayburt University, Faculty of Education, Department of Primary Education, Bayburt, Türkiye
[2]Bayburt University, Faculty of Education, Department of Mathematics and Science Education, Bayburt, Türkiye
[3]Bayburt University, Faculty of Education, Department of Educational Sciences, Bayburt, Türkiye

ARTICLE HISTORY

Received: June 08, 2023
Revised: July 28, 2023
Accepted: Oct. 18, 2023

Keywords:

Self-regulation,
Middle school students,
Scale development.

Abstract: This study was conducted to develop a valid and reliable measurement tool that can identify middle school students' self-regulation skills. Firstly, the literature was utilized in the development of the measurement tool. The form was finalized with the support of the opinions of different experts and a trial application. Then, the form was administered to 341 middle school students to determine its construct validity, and exploratory factor analysis (EFA) was performed on the collected data. Then, the form was administered to 341 middle school students to determine its construct validity, and exploratory factor analysis (EFA) was performed on the collected data. As a result of the study, it was determined that the scale consisted of 9 items and two sub-dimensions, namely "forethought" and "volitional control and self-reflection." In order to reveal the reliability of the scale, internal consistency, two-half test analyses, and composite reliability (CR) were used, and it was determined that the values found were .74 and above. In order to reveal the accuracy of the obtained structure, data were collected from 218 middle school students in a different province, and confirmatory factor analysis (CFA) was performed with these data. As a result of the analysis, the two-factor structure of the scale was confirmed. When the findings are examined, it can be said that the scale is a reliable and valid tool that can be applied to determine the self-regulation skills of students studying at the middle school level.

1. INTRODUCTION

Self-regulation has become an important topic in educational research in recent years, and research has generally focused on self-regulated learning (SRL) (Dever et al., 2023; Pijeira-Díaz, 2023). Therefore, SRL has been significantly examined in educational research. Although research on SRL has made a significant contribution to educational research, research on the process of self-regulation has remained limited (Öztürk, 2020). Research on the self-regulation process is important in evaluating the quality of student's own learning and the process of performing a task (Sökmen et al., 2023). In this context, it can be said that it is necessary to conduct research to examine the self-regulation process. The literature shows that the studies on self-regulation were mostly qualitative, and quantitative studies, including experimental and relational studies were insufficient in number (Öztürk, 2020). The reason for the low number

*CONTACT: Ismail SARIKAYA ✉ ismailsarikaya@bayburt.edu.tr Bayburt University, Faculty of Education, Department of Primary Education, Bayburt, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

of quantitative studies may be the lack of a scale that can determine students' self-regulation skills. This study was conducted to develop a valid and reliable scale to determine the self-regulation skills of middle school students. The research is expected to encourage the conduct of quantitative research on self-regulation skills.

1.1. Theoretical Framework

1.1.1. *Social cognitive theory*

In the literature, self-regulation is based on many different theories. Among these theories, the most accepted one for self-regulation is the social cognitive theory (Zimmerman, 2000). According to the social cognitive theory, the learner is at the center of learning (Zimmerman, 2000). According to this theory, learning takes place through the interaction of cognitive, environmental, and behavioral factors (Bandura, 1986). In other words, the events that the individual sees around him/her in the learning process, his/her own experiences, and his/her cognitive characteristics are effective (Bandura, 1986). Self-efficacy, self-regulation, and motivation are important concepts in social cognitive theory. This study focuses on the concept of self-regulation.

1.1.2. *Self-regulation*

The concept of self-regulation was first defined by Albert Bandura as an individual's setting goals, using resources to achieve them, and making the necessary effort (Bandura, 2002). Zimmerman (2000), who based self-regulation on social cognitive theory, defined self-regulation as planned and cyclical thoughts, feelings, and behaviors developed by the individual to achieve personal goals. Based on this definition, Zimmerman (2000) considered the self-regulation model in three dimensions: forethought, volitional control and self-reflection.

Forethought refers to the actions to be taken toward the task before starting a task (Sakız & Yetkin-Özdemir, 2014). In the forethought phase, the students analyze what is expected of them. Goal setting and strategic planning have a critical role in this phase (Zimmerman, 2000). In goal setting, the individual plans by deciding how to perform and the expectations at the end of the performance (Sakız & Yetkin-Özdemir, 2014). In this phase, the individual determines the strategies to be used in the process of performing the task before starting the task (Flower & Hayes, 1981; Pintrich, 2000; Wong et al., 2021; Zimmerman, 2000; Zimmerman & Risemberg, 1997). In strategic planning, the individual decides on the strategy necessary for the best performance. At this stage, the student can create and maintain the plan (Sakız & Yetkin-Özdemir, 2014).

Volitional control refers to the actions to be taken towards the task in the process of performing a task (Sakız & Yetkin-Özdemir, 2014). In the volitional control phase, the students can focus their attention to perform the task expected of them, take actions that will help him/her complete the task, and make the necessary changes in their plan (Zimmerman, 2000). In this stage, the individuals can keep distractions under control by focusing their attention on the subject of study. Thus, they can focus on the task (Sakız & Yetkin-Özdemir, 2014). In addition, individuals can control their time in the process and can make some changes if necessary (Öztürk & Ada, 2023). In the volitional control phase, students can control the strategies they have determined in the forethought phase and re-determine strategies if necessary (Sakız & Yetkin-Özdemir, 2014).

Self-reflection refers to the actions to be taken toward the task after performing a task (Sakız & Yetkin-Özdemir, 2014). In the self-reflection phase, the students complete the task and evaluate their performance on the completed task (Zimmerman, 2000). In this phase, the individual can check whether he/she has achieved the goals he/she has set, evaluate his/her performance in the process of performing the task, and identify his/her deficiencies (Zimmerman, 2000). In the self-reflection phase, individuals can perform the task by checking the strategies they determined in the forethought phase and, if necessary, by determining a new strategy (Öztürk & Ada, 2023).

In this study, Zimmerman's (2000) cyclical model of self-regulation was used. Although the model consists of three stages, it is very difficult to distinguish between skills that occur during the task process and skills that occur after the task. For example, controlling the student's goal is an action that occurs both in the process and in the outcome. Similarly, motivation is both a process and an outcome action. Therefore, in this study, a two-dimensional model was designed by considering volitional control and self-reflection together.

1.1.3. *The present study*

The literature review shows that different scales related to self-regulation have been developed. Eryılmaz and Mammadov (2017) developed a 47-item SRL scale based on Zimmerman's model as a result of their research with high school students aged 14-19. Another 16-item measurement tool is the perceived self-regulation scale (Arslan & Gelişli, 2015). Kröner et al. (2017) conducted a German adaptation study of the academic self-regulation questionnaire. The adaptation study of the 28-item scale developed by Steinbach and Stoeger (2018) to reveal teachers' attitudes toward SRL was carried out by Sarikaya and Sökmen (2021). Upon examination of these scales, it becomes evident that some of them are not suitable for middle school students due to the differences in the age groups of the sample in which they were developed, and some of them have weak usefulness due to the excessive number of items, and there are some points that do not match with the theoretical ground. It can be seen that the behaviors emphasized in some items in the scales are not related to self-regulation skills and that the items are not similar to the self-regulation behaviors mentioned in the existing theories. Since there is no general self-regulation scale developed directly for middle school students and the existing scales are not suitable in terms of item and structure, it was necessary to develop this scale. The current study aims to provide the literature with a measurement tool that is theoretically based on solid ground, can reveal self-regulation skills in middle school students in a valid and reliable way, and has high usability.

2. METHOD

In this section, information about the methodology of the study, such as the research model, the study group, the preparation of the trial form of the scale, the process, and data analysis are included.

2.1. Research Method

This study is a scale development study conducted in accordance with the survey model. The survey model is a type of research that is used to collect information on characteristics such as attitudes and behaviors from a large sample group. In this model, a general situation is presented (Fraenkel et al., 2015; McMillan & Schumacher, 2014). In this study, the steps of developing a Likert-type measurement tool were followed in order to determine the general self-regulation skills of middle school students. The study includes two main stages. In the first part, EFA was conducted to determine the structure of the scale; in the second part, CFA was conducted to verify the existing structure. Thus, the validity and reliability studies of the developed scale were reported in a phased manner.

2.2. Study Group

The participants for this study were selected using the cluster random sampling, which is a probability sampling technique. This sampling method involves selecting the sampling unit or units deemed necessary in the research and randomly obtaining participants from this group (Akarsu, 2015). This method is the best way designed to obtain a representative sample of the population of interest (Fraenkel et al., 2015). In this study, the simple random sampling method was used in order to increase the representative capacity of the population. The data of the study were obtained from two different provinces, one in Eastern Anatolia and the other in the Eastern Black Sea Region. First of all, five of the middle schools in the province were selected by random method. Then, the scale was applied to groups of voluntary participants.

The study was based on voluntary participation. The participants of the first phase of the study EFA consisted of 352 middle school students studying in the central district of a province in the Eastern Anatolia Region. All participants were Turkish and enrolled in public middle schools. The scale form belonging to six students, most of whom were found to be left blank in the pre-analysis checks, was excluded from the study; the analyses were carried out on 346 student data. In order to ensure representativeness in the study, an equal or close number of data were collected from all grade levels. Among the participants of the first phase, 85 were 5th-grade students, 86 were 6th-grade students, 86 were 7th-grade students, and 89 were 8th-grade students. Of the participants, 180 were girls (52%), and 166 were boys (48%).

The participants of the second stage of the study CFA consisted of 218 middle school students studying in the central district of a province located in the Eastern Black Sea Region. All participants were Turkish and enrolled in public middle schools. In the second phase of the study, an equal or similar number of students from each grade level were included in the study. Among the participants, 54 were 5th-grade students, 56 were 6th-grade students, 55 were 7th-grade students, and 53 were 8th-grade students. Of the participants, 111 were female (51%), and 107 were male (49%). Field (2013) states that in factor analysis studies, 100 people are considered poor, 200 people are considered average, 300 people are considered good, 500 people are considered very good, and 1000 people are considered excellent in terms of sample size. Considering the number of participants in the study, it can be concluded that the sample sizes are good and average in terms of meeting the requirements for factor analysis.

2.3. Preparation of the Trial Form of the Scale and Process

Firstly, ethics committee permissions and application permissions required for the study were obtained. In order to develop a measurement tool to determine the general self-regulation skills of middle school students, a trial form of the scale was first developed. The process of preparing the trial form involved several stages recommended in the literature, including item writing, seeking expert opinions, and conducting a pre-test (Tavşancıl, 2014). In order to write the scale items, firstly, a literature review was conducted, and the scales and studies in the literature were utilized (Arslan & Gelişli, 2015; Bandura, 1991, 2002; Diehl et al., 2006; Eryılmaz & Mammadov, 2017; Graham et al., 2005; Kröner et al., 2017; Pintrich & De Groot, 1990; Sarikaya & Sökmen, 2021; Sarikaya & Yılar, 2021; Schunk, 2005; Schunk & Zimmerman, 1994; Zimmerman, 2000, 2002). Accordingly, a trial form consisting of 32 items was developed.

It is recommended to consult expert opinion to ensure content and face validity before the pilot study (Kline, 1994). For this reason, the items were presented to three expert lecturers working on self-regulation, one expert lecturer in the field of measurement and evaluation, and two expert lecturers in the field of mother tongue education for comprehensibility. A form with options of "appropriate" and "inappropriate" was designed for each item to collect the experts' opinions on item suitability. The items were evaluated according to their ability to measure self-regulation skills, comprehensibility of the item, and language appropriateness. There was also a section in the form for the items that the experts wanted to add. As Veneziano and Hooper (1997) emphasized content validity was calculated as the ratio of the number of experts who expressed an appropriate opinion for the items to the total number of experts minus one. Items with content validity values below .80 were removed from the scale form; three new items were added to the form in line with the suggestions of the experts. Necessary revisions were made in terms of content and appearance, and the new items were presented to the experts for further evaluation, finalizing the scale. Through these processes, seven items were removed from the form, and resulting in a 28-item form for item analysis. In the final form, there are 25 positive and three negative items in the scale form to assess agreement with the items, a Likert-type five-point rating response category of "never (1)", "rarely (2)", "sometimes (3)", "usually (4)" and "always (5)" was used. In this form, the scale was tested for comprehensibility by applying it to 20 middle school students, and in line with the suggestions, minor changes were made on three items, and it was submitted to the experts again. As a result of the corrections, the form consisting of 28 items was finalized.

After these stages, validity and reliability studies of the revised form were conducted. First, the trial form was applied, and EFA was conducted. In addition, criterion-related validity was used within the scope of validity processes. In this context, the perceived self-regulation scale developed by Arslan and Gelişli (2015) was chosen as a criterion. Then, the revised form was applied, and CFA was conducted. It took approximately 10 minutes to complete the measurement tool.

2.4. Data Analysis

Before conducting the EFA, data preprocessing steps were implemented to ensure data quality. Extreme values, missing data, and incorrect values were addressed. To begin with, it is recommended that the number of cells left blank in the dataset should not exceed 2% of the total dataset. In this study, six participants' data were removed from the dataset as they contained extreme values or incorrect entries (Kalton & Kaspyzyk, 1986). The remaining missing data were examined, and it was observed that the number of cells left blank did not exceed 2% in the total data set; a total of seven cells were left blank in the entire data set. The missing data were filled with FIML method (Enders & Bandalos, 2001). For each student data set, kurtosis and skewness values and box-and-whisker plots were also analyzed. Then, the data set was analyzed in terms of multivariate outliers with the help of Mahalanobis distance. Five individuals who were significant at the $\alpha = 0.001$ level were excluded from the data set. All analyses were conducted with 341 student data. TV, VIF and CI index values were analyzed to determine whether there is a multicollinearity problem. As a result of the process, it was seen that there was no multicollinearity problem in the data set. Mardia's (1970) skewness value was used to determine whether the obtained data set meets the multivariate normal distribution. Uysal and Kılıç (2022) suggest the use of Mardia's skewness coefficient as a multivariate normality determination technique in terms of both power and type 1 error. Accordingly, it was determined that the data set did not exhibit a multivariate normal distribution (Mardia's skewness coefficient=109.01, $p<0.05$). It was also found that the skewness coefficients of the variables ranged from 0.100 (item 28) to -1.136 (item 27) and the kurtosis coefficient ranged from 0.004 (item 8) to -1.553 (item 28). Since the data set does not show a multivariate normal distribution, unweighted least squares (ULS), which is robust to violating this assumption of the analysis, was used as a factor extraction method (Brown & Moore, 2012).

Item correlation matrix, R-matrix determinant coefficients, and anti-image values were also examined for the suitability of the data set. Item correlation values are expected to be no lower than .30 and no higher than .90 (Field, 2013). It was determined that each item in the data set met this condition. Although the multicollinearity problem is not a major problem, it can be perceived as a problem in the case of extremely highly correlated items. In this case, the R-matrix determinant coefficient should be checked, and this value should be greater than 0.00001 (Field, 2013). In the study, it was observed that the values examined were greater than 0.00001. The values in the anti-image table are recommended to be greater than .50 (Field, 2013). The values obtained for the study data were found between .57 and .82. These findings show that the data are suitable for analysis. In addition, Kaiser-Mayer Olkin (KMO) value was examined for the adequacy of the sample size. The KMO value was calculated as .830. According to

Pallant (2001), a KMO value of 0.60 and above is considered sufficient. In addition, Bartlett's test of sphericity was significant ($X^2_{(36)}$=697,687; p=0.00). So, data driven correlation matrix is statistically significantly differ from identity matrix. In this direction, it can be said that the data obtained with the trial form are suitable for factor analysis. The distribution of items across factors and the loading values of items were analyzed. It is recommended that the difference between two loading values should be at least 0.10 to avoid overlapping items (Büyüköztürk, 2007). Overall, based on these analyses, the data were found to be suitable for further factor analysis, and overlapping items were identified and addressed. The relationship between the scale developed within the scope of this study and the perceived self-regulation scale (Arslan & Gelişli, 2015) selected as a criterion was examined with the Pearson correlation coefficient.

In order to determine the construct validity of the scale, factor analysis was performed with Direct Oblimin rotation. Because in cases where there is a relationship between sub-dimensions, it is recommended to use oblique methods (Kılıç, 2022). Scree Plot, optimal parallel analysis (Patil et al., 2017), was used to determine the number of factors. HULL analysis was not used to determine the number of factors. Because Kılıç (2022) states that as the correlation between dimensions increases, the success of HULL analysis in estimating the number of dimensions decreases. For item validity, corrected item-total correlations were checked and item factor loadings and common factor variance were taken into consideration. Internal consistency reliability was assessed using Cronbach's alpha coefficient and Mcdonald's omega coefficient. This coefficient measures the consistency or reliability of the scale as a whole and its sub-dimensions. Furthermore, CR were examined to assess the reliability of the scale. All of these analyses were performed using the SPSS 24.0 software package. CR calculated with Excel. Mardia'skewness coefficient was calculate with Past Istatistic 4.12.

Then, CFA was conducted to determine the model fit of the scale. Data set was analyzed in terms of multivariate outliers with the help of Mahalanobis distance. As a result of the analysis, no multivariate outliers were observed. TV, VIF and CI index values were analyzed to determine whether there is a multicollinearity problem. As a result of the process, it was seen that there was no multicollinearity problem in the data set. As a result of the analysis, it was seen that the data set met the assumption of multivariate normal distribution. For construct fit, X^2/df, RMSEA, SRMR, RMR, CFI, RFI, NNFI, NFI, and IFI fit indices were checked. In addition, R^2 values for each item were reported. The analyses were conducted with LISREL 8.80 package program.

3. FINDINGS

3.1. EFA Findings

In the first step of the analysis, the item-total correlation values were examined, and the items that were found to be insufficient were removed from the analysis in order (Items 28, 7, 17, 14, 10, 27, 26, 8, 11). The analysis was repeated after each item's removal.

According to the results of the factor analysis conducted with 19 items, a four-factor structure with eigenvalues above 1, at least two items, variance explained by more than 5%, and 48.72% of the total variance was obtained. The ratios of each item explaining the variance in a common factor together were examined, and it was observed that these values ranged between .31 and .60. Kalaycı (2009) stated that removing the items with low factor loadings from the analysis will increase the total variance explained. Accordingly, items 2, 12, and 20 in the trial form with a common variance value of less than .40 were removed from the analysis, and the factor analysis was conducted again. As a result of the analysis with 16 items, a three-factor structure explaining 46.88% of the total variance was reached. The result of the analysis showed that the factor loadings of some items was less than .40. For this reason, Items 23, 16, and 15 were removed from the analysis, respectively. The analysis was repeated after each item's removal. In the final analysis with 13 items, a three-factor structure explaining 51.33 of the total variances emerged. In this model, the factor variances of the items are between .40 and .59.

EFA was conducted on the new scale form obtained. Two items that were found to be overlapping items were removed from the form. After the items were removed, the explained variance decreased to 47%. The analysis was repeated and it was determined that the common variance value of one item decreased to .29. For this reason, the related item was removed from the analysis. The analysis was repeated and it was seen that the common factor variance value of one item was .35. Therefore, the related item was removed from the analysis. As a result of this process, nine items remained on the scale. All of the remaining items are positive.

For the nine-item version of the form, the item correlation matrix, R-matrix determinant coefficients, and anti-image values were re-examined before EFA. As a result of this process, it was observed that the relevant values were within the ideal range. It was observed that the correlation values were moderate, determinant coefficients were greater than 0.01, anti-image covariance values were between .57 and .77, and anti-image correlation values were between .80 and .89.

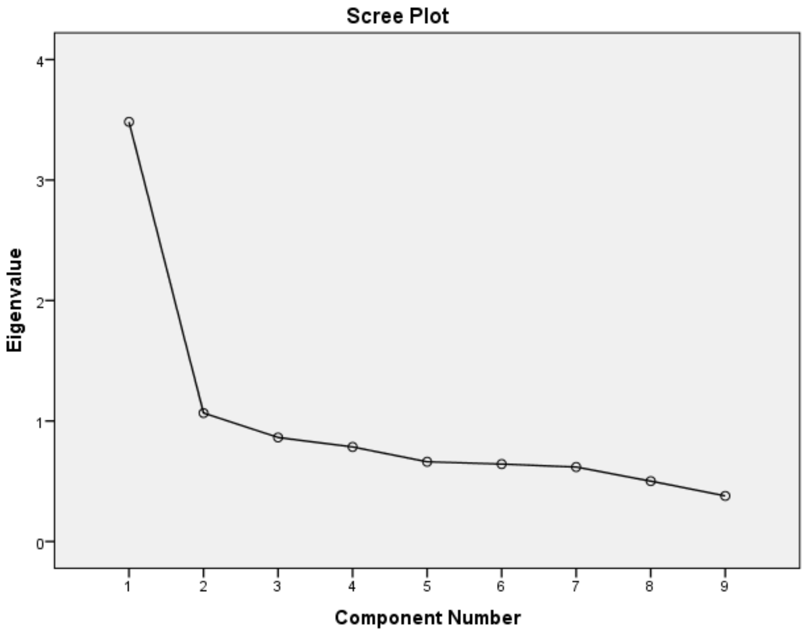
EFA was conducted again with the remaining nine items in the trial form, and the factor properties obtained are presented in Table 1.

Table 1. *Characteristics of the factors.*

Factor	Eigenvalue	Variance (%)	Total Variance (%)
1. Forethought	3.48	38.70	38.70
2. Volitional control and self-reflection	1.07	11.85	50.55

As seen in Table 1, according to the EFA results, a two-dimensional structure explaining 50.55% of the total variance, having at least 1% eigenvalue, at least four items, and at least 12% variance, was obtained. The Scree Plot showing the number of factors is shown in Figure 1.

Figure 1. *Scree plot.*



The Scree Plot shown in Figure 1 reveals a two-factor structure. To determine factor number optimal parallel analysis was used. Optimal parallel analysis findings reveal that Optimal parallel analysis findings reveal that it is a one-factor structure (3.48>1.417 for one factor; but

1.07<1.33). In this case, it is recommended to consult expert opinions. In order to determine the number of factors, the opinions of two faculty members who have international competence in the field of self-regulation and who also have self-regulation theory were consulted. Both experts who examined the items and factors approved the two-factor structure of the scale. One of the experts even presented the fact that the first four items targeted cognitive and the last five items targeted affective features as evidence. Therefore, the two-factor structure of the scale was analyzed. The results of the rotated component analysis (item number, item content, common factor variance of the items, factors, and factor loadings of the items under the factors) are presented in Table 2.

Table 2. *Rotated factor analysis results.*

Factor Name	Item	Common Factor Variance (Extraction)	Factor 1	Factor 2
Forethought	1	.57	.76	
	2	.45	.67	
	3	.62	.82	
	4	.42	.49	
Volitional control and self-reflection	5	.48		.64
	6	.53		.57
	7	.50		.64
	8	.49		.74
	9	.49		.74

Table 2 show that the ratios of each item explaining the variance in a common factor together were examined, and it was observed that these values ranged between .45 and .62. In the two-factor structure obtained, the first factor is the “forethought” sub-dimension consisting of Items 1-4. The second one is the “volitional control and self-reflection” sub-dimension consisting of Items 5-9. While naming the dimensions, the structure that the items aimed to measure was taken into consideration, and expert opinions were consulted. The loading values of the items in the “forethought” factor ranged between .49 and .82; the loading values of the items in the “volitional control and self-reflection” factor ranged between .57 and .74.

To determine the relationship between the factors of the scale, an analysis of the correlation between the factors was conducted. The resulting data are presented in Table 3.

Table 3. *Correlation between factors.*

Factors	r
1. Forethought	1.00
2. Volitional control and self-reflection	.533**

$^{**}p$<.01

Table 3 shows the correlation between sub-dimensions. The analysis indicates a positive, and significant relationship between them (r=.533, *p*=.000). This finding shows that there is no multicollinearity problem between the sub-dimensions of the scale and that they measure a separate characteristic. It was determined that there is a positive, high and significant relationship between the perceived self-regulation scale developed by Arslan and Gelişli (2015) for criterion-related validity and the scale developed within the scope of the study (r=.724, *p*=.000). Internal consistency Cronbach’s Alpha and Mcdonald Omega reliabilities for the whole scale and for its sub-dimensions CR were calculated on the same sample. The coefficients are shown in Table 4.

Table 4. *Reliability coefficients.*

	Cronbach's Alpha	Mcdonald Omega	CR
Forethought	.78	.70	.77
Volitional control and self-reflection	.83	.73	.80
Overall Scale	.90	.79	

The Cronbach Alpha internal consistency coefficient of the overall scale was calculated as .90, Mcdonald Omega as .79. Considering these values, it can be considered that the scale has the necessary reliability. The fact that the reliability coefficients of the sub-dimensions are above .70 provides evidence that the scale will provide reliable measurements. Table 4 shows that CR values for all factors are higher than .70.

3.2. CFA Findings

CFA was conducted to evaluate the adequacy of the structure obtained from the EFA. As a result of CFA, it was determined that the chi-square fit index of the structure consisting of nine items and two factors was significant ($X^2_{(26)}$=44.20, p=.014). In this case, it is recommended to look at other fit indices (Table 5). R^2 values of the items are also presented in Table 5.

Table 5. *Fit index values.*

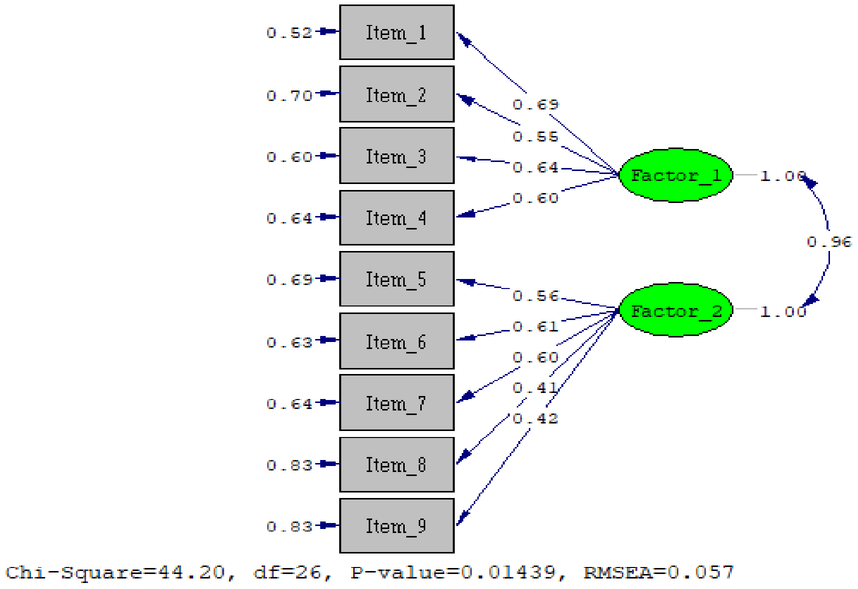
Indexes	Perfect Fit Criterion	Acceptable Fit Criterion*	Finding	Conclusion	Item	R^2
X^2 /df	0-2.5	2.5-3	1.7	Perfect	1	.48
SRMR	≤.05	≤.08	.045	Perfect	2	.30
IFI	≥.95	≥.90	.98	Perfect	3	.40
NNFI	≥.95	≥.90	.97	Perfect	4	.36
CFI	≥.95	≥.90	.97	Perfect	5	.31
RMSEA	≤.05	≤.08	.057	Acceptable	6	.37
RMR	≤.05	≤.08	.071	Acceptable	7	.36
RFI	≥.95	≥.90	.92	Acceptable	8	.17
NFI	≥.95	≥.90	.94	Acceptable	9	.17

*(Schumacker & Lomax, 2004)

Upon reviewing the fit index values presented in Table 5, it is observed that X^2 /df, SRMR, NNFI, CFI, and IFI values have excellent fitness levels; RMSEA, RMR, NFI, and RFI values have an acceptable fit level. Within the scope of the study, R^2 values for each item were also examined. It is seen that the R^2 values of the items are between .17 and .48. The R^2 values of the items are an indicator of the variance explanation rates of the items. In order to check the construct fit, the diagram containing the item-total correlation values and the t-value diagram were also examined. The path diagram for the first level CFA is shown in Figure 2.

Figure 2 depicts the item-total correlation values, which range between .41 and .69. These values indicate the correlation between each variable and the latent variable. A higher correlation coefficient suggests a greater degree of the variable's contribution in explaining the latent variable. Therefore, based on these findings, it can be inferred that the items in the scale are effective in distinguishing the trait to be measured and that they capture similar behaviors. Furthermore, a thorough examination was conducted to determine the significance of the analysis values for each item. It was found that the t-values ranged between 5.62 and 10.39. Importantly, all items reached significance at the p<.01 level. Thus, it can be confidently stated that the items within the scale are designed to measure distinct characteristics.

Figure 2. *Path diagram.*



4. DISCUSSION and CONCLUSION

The findings of this study, which aimed to develop a valid and reliable measurement tool that can measure the self-regulation skills of middle school students, supported the literature and provided some original results to the literature. The results of the study showed that the self-regulation scale consists of two dimensions: “forethought” and “volitional control and self-reflection”. In this respect, it can be said that the research supports the literature. The unique result obtained in the study was that the items related to volitional control and self-reflection skills were collected in the same dimension. This may be due to the student’s inability to fully distinguish between process and outcome self-regulation skills. In other words, students may not be able to distinguish between process and outcome evaluations. As a result of the research, a valid and reliable measurement tool was developed to measure the self-regulation skills of middle school students. English and Turkish versions of the scale form are presented in the Appendix.

Within the scope of the study, an item pool was first created. In order to write the scale items, a literature review was conducted, and the scales and studies in the literature were utilized (Arslan & Gelişli, 2015; Bandura, 1991, 2002; Diehl et al., 2006; Eryılmaz & Mammadov, 2017; Graham et al., 2005; Kröner et al., 2017; Pintrich & De Groot, 1990; Sarikaya & Sökmen, 2021; Sarikaya & Yılar, 2021; Schunk, 2005; Schunk & Zimmerman, 1994; Zimmerman, 2000, 2002). The items were then submitted to expert opinion in terms of content, language, and face validity. Then, the suitability of the data set for analysis was tested, and extreme data were removed from the data set. The scale was administered to 341 middle school students for EFA and 218 middle school students for CFA. Comrey and Lee (1992) and Field (2013) state that 100 is a poor sample size, 200 is a fair sample size, 300 is a good sample size, 500 is a very good sample size, and 1000 is an excellent sample size in factor analysis studies. Considering the number of participants in the study, it can be concluded that the sample sizes are good and average. Before EFA, the item correlation matrix, R-matrix determinant coefficients, and anti-image values were analyzed. Field (2013) states that item correlation values between .30 and .90 are a criterion for the suitability of the data set for analysis. The findings of the present study reveal that the item correlation values are within the aforementioned range. Field states that multicollinearity is not a major problem in scale development but suggests that this situation should be checked. For this reason, R-matrix determinant coefficients were checked, and it was seen that these values were greater than 0.01. This shows that there is no multicollinearity among the items. In addition, the values in the anti-image table were found to be at the ideal level (>.50) (Durmuş et al., 2013). The KMO test result (.830), which shows the adequacy of the sample size, reveals that the data set is large enough for analysis. Pallant (2001) states that this value should be .60 and above. The result of Barlett's test of sphericity was also significant, indicating that the data set was suitable for factor analysis. Consequently, factor analysis was performed using direct oblimin rotation to assess the construct validity of the scale based on the obtained data.

In the EFA findings, the rates of each item explaining the variance in a factor loading together were examined. Kalaycı (2009) stated that removing variables with low factor loadings from the analysis will increase the overall explained variance. Accordingly, the items in the trial form with a factor loading less than .40 were removed from the analysis, and the factor analysis was conducted again. The extraction values of the items (.419-.622) indicate a satisfactory level. After this stage, the item factor loading values and the distribution of the items to the factors were analyzed. In item factor loading values, the common factor loading value is in favor of .30 and above; however, weights of .50 and above are considered quite good (Field, 2013; Kalaycı, 2009; Tavşancıl, 2014). In this direction, it was paid attention that the item factor loading value was .50 and above (.49-.82). Therefore, it can be concluded that the factor loading values of the scale items are highly satisfactory. Within the scope of the study, items that were found to be overlapping items were removed from the form. As a result of all these analyses, EFA was repeated, and a structure consisting of nine items and two sub-dimensions was obtained, explaining approximately 51% of the total variance. Kline (1994) emphasizes that the variance explained by the measurement tool should be at least 40%; Henson and Roberts (2006) emphasize that it should be 52%. Given these criteria, the variance explained in this study can be considered sufficient. A high, positive and significant relationship was found between the scale's perceived self-regulation scale (Arslan & Gelişli, 2015) used for criterion validity. This result provides evidence for the validity of the general self-regulation scale developed within the scope of the study.

The correlation value between the factors of the scale was analyzed (r=.533, *p*=000). A correlation coefficient of .90 and above between sub-dimensions is not recommended because it may indicate a multicollinearity problem (Field, 2013; Pallant, 2001). These data show that there are significant relationships between the sub-dimensions of the scale and that there is no multicollinearity problem. In addition, based on the findings, it can be said that each sub-factor measures a separate feature. The reliability analysis findings of the scale also provide evidence that the scale will provide reliable measurements (Fornell & Larcker, 1981).

The fit of the structure obtained from EFA was assessed using CFA. The chi-square fit index was found to be significant, indicating the need to examine other fit indices. The X^2/df value, which was calculated to assess model fit, was less than 2.5. According to Kline (2005), a value below 2.5 suggests excellent fit for the structure. Furthermore, Schumacker and Lomax (2004) and Kline suggested that the SRMR, NNFI, CFI, and IFI values demonstrate excellent fit, while the RMSEA, RMR, NFI, and RFI values indicate an acceptable fit. The R^2 values of the scale items provide insight into the contributions of each item to the model. Additionally, the item-total correlation values were found to be .40 or above. Büyüköztürk (2007) suggests that these values should be at least .30 or higher in order to effectively distinguish the feature being measured. Therefore, it can be concluded that the item-total correlation values indicate a good fit for the model. Similarly, the t values of the items confirm that each item aims to measure a distinct characteristic.

The scale has two sub-dimensions: "forethought" and "volitional control and self-reflection". While naming the sub-dimensions, the structure and characteristics of the items were taken into consideration as suggested by Kalaycı (2009). The first item of the "forethought" sub-dimension is related to goal setting. Many researchers emphasize that goal setting is a critical self-regulation skill and takes place in the forethought stage (Flower & Hayes, 1981; Pintrich, 2000; Wong et al., 2021; Zimmerman, 2000; Zimmerman & Risemberg, 1997). Schunk (1996, 2003) considers goal setting as an important factor that initiates self-regulation processes. The second item is related to strategizing. Wood and Bandura (1989) state that strategy development exists both at the beginning (forethought) and at the self-evaluation stage of the self-regulation process. There are researchers who state that strategy development is a critical self-regulation behavior (Bereiter & Scardamalia, 1983; De Smedt et al., 2018; Zimmerman, 2000). The third item involves creating a study plan. Planning skill is also related to strategy development skill. Researchers state that individuals with high levels of self-regulation skills plan frequently, check their plans, and revise their plans when necessary (Graham & Harris, 2005; Zimmerman & Kitsantas, 2007). Wong et al. (2021) report that planning is related to self-regulation skills and is an important factor in developing self-regulation skills. The last item of the forethought factor is related to the behavior of sticking to the plan. This is related to individuals' maintaining their motivation and self-efficacy beliefs (Graham et al., 1998; Raphael et al., 1988). The effect of motivation and self-efficacy beliefs on the self-regulation process is undeniable (Agustiani et al., 2016). As a matter of fact, Bandura (1993) and Schunk (2003) report that one of the most important sensory characteristics in the self-regulation process is self-efficacy belief. Malmivuori (2006) and Su et al. (2018) state that individuals with strong self-efficacy use self-regulation skills more actively. In this context, it can be stated that each of the scale items in the "forethought" sub-dimension requires self-regulation skills and are critical skills for determining general self-regulation skills. In this direction, it can be stated that the current study supports the literature.

When the items in the "volitional control and self-reflection" dimension are examined, it can be seen that there are items that include sensory characteristics such as motivation, attention, focus and time control. The first item is related to the willingness to learn. This item focuses on the individual's motivational processes. Researchers state that there is a strong relationship between self-regulation and motivation (Bandura, 1993; Op't Eynde et al., 2007; Li et al., 2022; Yu et al., 2022). The second item is about paying attention, and the third item is about focusing behaviors. Hanif et al. (2012) state that attention regulation includes self-regulation. Studies reveal the positive effect of attention and focus on self-regulation and report that individuals with high self-regulation are able to control attention and focus on their work (Berger et al., 2007; Koopmann et al., 2019; Posner & Rothbart, 2009; Sassenberg & Woltin, 2008; Turcotte et al., 2022). The last two items of the "volitional control and self-reflection" sub-dimension are related to time control and time rescheduling. Time control is a critical self-regulation behavior. Pajares and Johnson (1994) and Zimmerman and Risemberg (1997) state that effective management and planning of time is included in the personal processes sub-dimension of self-regulation. Indeed, individuals with high self-regulation who set goals and develop strategies are expected to be successful in time management (Sarikaya & Yılar, 2021). Harris and Graham (1996) also report that using time effectively and avoiding wasting time are self-regulation skills related to temporal factors. As can be seen, it is clear that each item that constitutes the dimension of "volitional control and self-reflection" is associated with self-regulation skills in the literature. In this direction, it can be stated that the findings support the literature.

4.1. Recommendations and Contributions to Education

As a result, it was determined that the scale developed was sufficiently reliable both on a general basis and on the basis of factors. It was concluded that the items of the scale could measure the trait that it aims to measure and can distinguish between individuals who have the trait targeted to be measured and individuals who do not have it. The scale can be applied to all middle school students studying at different grade levels. The scale can be used in different studies to contribute to its validity and reliability. In addition, a primary school form of the scale can also be created based on the basic self-regulation skills included in the scale items. For this purpose, validity and reliability studies can be conducted again, and the compatibility between the two scales can be revealed.

Acknowledgments

This work was financially supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant number 121K859.

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**: Bayburt University, E-51694156-050.99-36868).

Authorship Contribution Statement

Ismail Sarikaya: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Mesut Ozturk**: Visualization, Software, Methodology, Supervision, and Validation. **Mustafa Ozgol:** Methodology, Supervision, and Validation.

Orcid

Ismail Sarikaya https://orcid.org/0000-0002-4870-8345
Mesut Ozturk https://orcid.org/0000-0002-2163-3769
Mustafa Ozgol https://orcid.org/0000-0002-9493-3455

REFERENCES

Agustiani, H., Cahyad, S., & Musa, M. (2016). Self-efficacy and self-regulated learning as predictors of students academic performance. *The Open Psychology Journal, 9*(1), 1-6. https://doi.org/10.2174/1874350101609010001

Akarsu, B. (2015). Hipotezlerin, değişkenlerin ve örneklemin belirlenmesi [Determining hypotheses, variables and sample]. M. Metin. (Eds.), In *Kuramdan uygulamaya eğitimde bilimsel araştırma yöntemleri [Scientific research methods in education from theory to practice]* (2nd ed., pp. 21-43). Pegem.

Arslan, S., & Gelişli, Y. (2015). Development of perceived self-regulation scale: Validity and reliability study. *Sakarya University Journal of Education, 5*(3), 67-74. http://dx.doi.org/10.19126/suje.91303

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Prentice-Hall.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248-287. https://doi.org/10.1016/0749-5978(91)90022-L

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117-148. https://doi.org/10.1207/s15326985ep2802_3

Bandura, A. (2002). Social cognitive theory in cultural context. *Applied Psychology: An International Review, 51*(2), 269-290. https://doi.org/10.1111/1464-0597.00092

Bereiter, C., & Scardamalia, M. (1983). Levels of inquiry in writing research. In P. Rosenthal, S. Walmsley & L. Tamor (Eds.), *Research in writing: Principles and methods* (pp. 3-25). Longman International.

Berger, A., Kofman, O., Livneh, U., & Henik, A. (2007). Multidisciplinary perspectives on attention and the development of self-regulation. *Progress in Neurobiology, 82*(5), 256-286. https://doi.org/10.1016/j.pneurobio.2007.06.004

Brown, T.A., & Moore, M.T. (2012). Confirmatory factor analysis. In R.H. Hoyle (Ed.), *Handbook of structural equation modeling* (*pp.* 361-379). Guilford.

Büyüköztürk, Ş. (2007). *Sosyal bilimler için veri analizi el kitabı [Manual of data analysis for social sciences].* Pegem.

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis.* Erlbaum.

De Smedt, F., Merchie, E., Barendse, M., Rosseel, Y., De Naeghel, J., & Van Keer, H. (2018). Cognitive and motivational challenges in writing: Studying the relation with writing performance across students' gender and achievement level. *Reading Research Quarterly, 53*(2), 249-272. https://doi.org/10.1002/rrq.193

Dever, D.A., Sonnenfeld, N.A., Wiedbusch, M.D., Schmorrow, S.G., Amon, M.J., & Azevedo, R. (2023). A complex systems approach to analyzing pedagogical agents' scaffolding of SRL within an intelligent tutoring system. *Metacognition and Learning*. https://doi.org/10.1007/s11409-023-09346-x

Diehl, M., Semegon, A.B., & Schwarzer, R. (2006). Assessing attention control in goal pursuit: A component of dispositional self-regulation. *Journal of Personality Assessment, 86*(3), 306-317. https://doi.org/10.1207/s15327752jpa8603_06

Durmuş, B., Yurtkoru, E.S., & Çinko, M. (2013). *Sosyal bilimlerde SPSS'le veri analizi [Data analysis with SPSS in social sciences]* (5th ed.). Beta.

Enders, C.K., & Bandalos, D.L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430-457. https://doi.org/10.1207/S15328007SEM0803_5

Eryılmaz, A., & Mammadov, M. (2017). Development of a self-regulatory learning measurement on the model of Zimmerman. *The Journal of International Educational Sciences, 4*(10), 79-93. https://doi.org/10.16991/INESJOURNAL.1360

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Flower, L., & Hayes, J.R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387. https://doi.org/10.2307/356600

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50.

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2015). *How to design and evaluate research in education* (9th ed.). Mc Graw Hill Education.

Graham, S., Harris, K.R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*(2), 207-241. https://doi.org/10.1016/j.cedpsych.2004.08.001

Graham, S., Harris, K.R., & Troia, G.A. (1998). Writing and self-regulation: Cases from the self-regulated strategy development model. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self regulated learning: From teaching to self reflective practices* (pp. 20-41). Guilford Press.

Hanif, A., Ferrey, A.E., Frischen, A., Pozzobon, K., Eastwood, J.D., Smilek, D., & Fenske, M.J. (2012). Manipulations of attention enhance self-regulation. *Acta Psychologica, 139*(1), 104-110. https://doi.org/10.1016/j.actpsy.2011.09.010

Harris, K.R., & Graham, S. (1996). *Making the writing process work: Strategies for composition and self-regulation.* Brookline Books.

Henson, R.K., & Roberts, J.K. (2006). Use of exploratory analysis in published research: Common errors and some comments on improved practice. *Educational and Psychological Measurement*, *66,* 393-416. http://dx.doi.org/10.1177/0013164405282485

Kalaycı, Ş. (2009). *SPSS uygulamalı çok değişkenli istatistik teknikleri [SPSS applied multivariate statistical techniques]* (4th ed.). Asil.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology, 12*, 1-16.

Kılıç, A.F. (2022). Deciding the number of dimensions in explanatory factor analysis: A brief overview of the methods. *Pamukkale University Journal of Social Sciences Institute, 51*(1-special issue), 305-318. https://doi.org/10.30794/pausbed.1095936

Kılıç, A.F. (2022). Ölçek geliştirme sürecinde açımlayıcı faktör analizi [Exploratory factor analysis in scale development process]. In M. Acar-Güvendir & Y. Özer-Özkan (Eds.), *Tüm yönleriyle ölçek geliştirme süreci [All aspects of the scale development process]* (*pp.* 69-129). Pegem.

Kline, P. (1994). *An easy guide to factor analysis*. Routledge.

Koopmann, J., Johnson, R.E., Wang, M., Lanaj, K., Wang, G., & Shi, J. (2019). A self-regulation perspective on how and when regulatory focus differentially relates to citizenship behaviors. *Journal of Applied Psychology*, *104*(5), 629-641. https://doi.org/10.1037/apl0000366

Kröner, J., Goussios, C., Schaitz, C., Streb, J., & Sosic-Vasic, Z. (2017). The construct validity of the German academic self-regulation questionnaire (SRQ-A) within primary and middle school children. *Frontiers in Psychology, 8*(1032), 1-13. https://doi.org/10.3389/fpsyg.2017.01032

Li, G., Luo, H., Lei, J., Xu, S., & Chen, T. (2022). Effects of first-time experiences and self-regulation on college students' online learning motivation: Based on a national survey during COVID-19. *Education Sciences, 12*(4), 245. https://doi.org/10.3390/educsci12040245

Malmivuori, M.L. (2006). Affect and self-regulation. *Educational Studies in Mathematics, 63*, 149-164. https://doi.org/10.1007/s10649-006-9022-8

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530. https://doi.org/10.2307/2334770

McMillan, J., & Schumacher, S. (2014). *Research in education: Evidence-based inquiry* (7th ed.). Pearson Education Limited.

Op 't Eynde, P., De Corte, E., & Verschaffel, L. (2007). Students' emotions: A key component of self-regulated learning? In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 185-204). Elseiver. https://doi.org/10.1016/B978-0-12-372545-5.X5000-X

Öztürk, M. (2020). The relationship between self-regulation and proportional reasoning: The mediating role of reflective thinking towards problem solving. *Education and Science, 45*(204), 143-155. https://doi.org/10.15390/EB.2020.8480

Öztürk, M. & Ada, K. (2023). Metacognition in problem solving and problem posing. In K. Özgen, T. Kar, S. Çenberci, Y. Zengin (Eds.) *Problem solving and problem posing in mathematics* (pp. 115-138). Pegem.

Pajares, F., & Johnson, M.J. (1994). Confidence and competence in writing: The role of self-efficacy, outcome expectancy, and apprehension. *Research in the Teaching of English, 28*(3), 313-331. https://www.jstor.org/stable/40171341

Pallant, J. (2001). *SPSS survival manual.* Open University Press.

Patil, V.H., Surendra, N., Singh, S.N., Mishra, S., & Donavan, D.T. (2017). *Parallel analysis engine to aid in determining number of factors to retain using R*. [Computer software]. https://analytics.gonzaga.edu/parallelengine/.

Pijeira-Díaz, H.J., van de Pol, J., Channa, F., & de Bruin, A. (2023). Scaffolding SRL from causal-relations texts: Diagramming and self-assessment to improve metacomprehension accuracy?. *Metacognition and Learning*, 1-28. https://doi.org/10.1007/s11409-023-09343-0

Pintrich, P.R. (2000). The role of goal orientation in self-regulated learning. *Handbook of Self-Regulation*, 451–502. https://doi.org/10.1016/b978-012109890-2/50043-3

Pintrich, P.R., & De Groot, E.V. (1990). Motivational and SRL components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33-40. https://doi.org/10.1037//0022-0663.82.1.33

Posner, M.I., & Rothbart, M.K. (2009). Toward a physical basis of attention and self regulation. *Physics of Life Reviews*, *6*(2), 103-120. https://doi.org/10.1016/ j.plrev.2009.02.001

Raphael, T.E., Kirschner, B.W., & Englert, C.S. (1988). Expository writing program: Making connections between reading and writing. *The Reading Teacher*, *41*(8), 790-795. https://www.jstor.org/stable/20199924

Sakız, G. & Yetkin-Özdemir, İ.E. (2014). Özdüzenleme ve özdüzenlemeli öğrenme: Kuramsal bakış [Self-regulation and SRL: A theoretical perspective]. In G. Sakız (Ed.) *Özdüzenleme: Öğrenmeden öğretime özdüzenleme davranışlarının gelişimi, stratejiler ve öneriler [Self-regulation: Development of self-regulation behaviors from learning to teaching, strategies and suggestions]*. (pp. 2-23). Nobel Academy.

Sarikaya, İ., & Sökmen, Y. (2021). Adaptation of the teacher attitudes towards SRL scale to Turkish and examining of primary school teachers' attitudes towards SRL. *Erzincan University Journal of Education Faculty*, *23*(1), 126-147. https://doi.org/10.17556/erziefd.730175

Sarikaya, İ., & Yılar, Ö. (2021) Exploring self-regulation skills in the context of peer assisted writing: Primary school students' sample. *Reading & Writing Quarterly*, *37*(6), 552-573. https://doi.org/10.1080/10573569.2020.1867677

Sassenberg, K., & Woltin, K.-A. (2008). Group-based self-regulation: The effects of regulatory focus. *European Review of Social Psychology*, *19*, 126-164. https://doi.org/10.1080/10463280802201894

Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling*. Psychology Press.

Schunk, D.H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, *33*(2), 359-382. https://doi.org/10.3102/00028312033002359

Schunk, D.H. (2003). Self-efficacy for reading and writing: Influence of modeling, goal setting, and self evaluation. *Reading & Writing Quarterly*, *19*(2), 159-172. https://doi.org/10.1080/10573560308219

Schunk, D.H. (2005). SRL: The educational legacy of Paul R. Pintrich. *Educational Psychologist*, *40*(2), 85-94. https://doi.org/10.1207/s15326985ep4002_3

Schunk, D.H., & Zimmerman, B.J. (1994). *Self-regulation of learning and performance: Issues and educational applications*. Lawrence Erlbaum.

Sökmen, Y., Taş, Y., & Sarikaya, İ. (2023). An evaluation of the studies on SRL in primary education: A bibliometric mapping analysis. *Psycho-Educational Research Reviews*, *12*(1), 321-337. https://doi.org/10.52963/PERR_Biruni_V12.N1.20

Steinbach, J., & Stoeger, H. (2018). Development of the teacher attitudes towards SRL scale. European Journal of Psychological Assessment, 34(3), 193-205. https://doi.org/10.1027/1015-5759/a000322

Su, Y., Zheng, C., Liang, J.C., & Tsai, C.C. (2018). Examining the relationship between English language learners' online self-regulation and their self-efficacy. *Australasian Journal of Educational Technology*, *34*(3), 105-121. https://doi.org/10.14742/ajet.3548

Tavşancıl, E. (2014). *Tutumların ölçülmesi ve SPSS ile veri analizi [Measuring attitudes and data analysis with SPSS] (*5th ed.). Nobel.

Turcotte, J., Lakatos, L., & Oddson, B. (2022). Self-regulation of attention may unify theories of mindfulness. Psychology of Consciousness: *Theory, Research, and Practice*. Advance online publication. https://doi.org/10.1037/cns0000310

Uysal, İ., & Kılıç, A.F. (2022). Çok değişkenli normallik: Testler ne kadar doğru ne kadar güçlü? [Multivariate normality: How accurate and powerful are the tests?] In F. Nayır & Ş. Poyrazlı (Ed.), *Eğitim Bilimlerinde Güncel Araştırmalar [Current Research in Educational Sciences]*. Anı.

Veneziano, L., & Hooper, J. (1997). A method for quantifying content validity of health related questionnaires. *American Journal of Health Behavior*, *21*(1), 67-72.

Wong, J., Baars, M., He, M., de Koning, B. B., & Paas, F. (2021). Facilitating goal setting and planning to enhance online self-regulation of learning. *Computers in Human Behavior*, *124*, 106913. https://doi.org/10.1016/j.chb.2021.106913

Wood, R., & Bandura, A. (1989). Social cognitive theory of organizational management. *Academy of Management Review*, *14*(3), 361-384. https://doi.org/10.2307/258173

Yu, J., Huang, C., He, T., Wang, X., & Zhang, L. (2022). Investigating students' emotional self-efficacy profiles and their relations to self-regulation, motivation, and academic performance in online learning contexts: A person-centered approach. *Education and Information Technologies.* Advance online publication. https://doi.org/10.1007/s10639-022-11099-0

Zimmerman, B.J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P.R. Pintrich & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). USA: Academic. http://dx.doi.org/10.1016/B978-012109890-2/50048-2

Zimmerman, B.J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practise*, *41*(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2

Zimmerman, B.J., & Kitsantas, A. (2007). A writer's discipline: The development of self-regulatory skill. In P.Boscolo & S. Hidi (Eds.), *Writing and motivation* (pp. 51-69). Elsevier.

Zimmerman, B.J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology*, *22*(1), 73-101. https://doi.org/10.1006/ceps.1997.0919

APPENDIX

Turkish and English versions of the scale

Self-Regulation Scale for Middle School Students

Items	**Never ever**	**Rarely**	**Sometimes**	**Usually**	**Always**
1. I set my goals before I start working.					
2. I strategize how I will learn before I start working.					
3. I create a work plan.					
4. I try to stick to my plan.					
5. I can make myself willing to learn.					
6. I focus my attention on the work.					
7. I focus on what I do.					
8. I check whether I have enough time.					
9. I can reschedule time according to the disruptions in the process.					

Ortaokul Öğrencileri İçin Öz-Düzenleme Ölçeği

Maddeler	**Hiçbir zaman**	**Ara sıra**	**Bazen**	**Genellikle**	**Her zaman**
1. Çalışmaya başlamadan önce hedeflerimi belirlerim.					
2. Çalışmaya başlamadan önce nasıl öğreneceğime ilişkin stratejiler oluştururum.					
3. Çalışma planı oluştururum.					
4. Planıma sadık kalmaya çalışırım.					
5. Kendimi öğrenmeye istekli hale getirebilirim.					
6. Dikkatimi çalışma konusuna veririm.					
7. Yaptığım işe odaklanırım.					
8. Yeterli zamanımın olup olmadığını kontrol ederim.					
9. Süreçte çıkan aksaklıklara göre zamanı yeniden planlayabilirim.					

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 741–764

https://doi.org/10.21449/ijate.1352605

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

An examplary application of mathematics teaching based on formative assessment

Hilal Ozcan[1,*], Aytac Kurtulus[2]

[1]Ministry of Education, Osmaniye, Türkiye
[2]Eskişehir Osmangazi University, Faculty of Education, Department of Educational Sciences, Eskişehir, Türkiye

ARTICLE HISTORY

Received: Aug. 31, 2023
Revised: Nov. 12, 2023
Accepted: Nov. 20, 2023

Keywords:
Mathematics teaching,
Formative assessment,
Mathematics achievement,
Circle and circular region.

Abstract: Mathematics, by nature, has a spiral structure. The fact that the students fail to acquire an achievement in the mathematics course negatively affects other achievements. Therefore, students' mathematics failure continues exponentially. In this context, it is important to identify and eliminate students’ weaknesses, if any, with the formative assessments in the learning process. The present study aimed to improve the success level of the students with the formative assessment-based teaching practice about circle and circular region in secondary school seventh-grade mathematics teaching and implement an exemplary application for formative assessment-based teaching.

Since the study sought to evaluate the learning process of the students and eliminate the weaknesses identified in the process, an action research design, one of the methods of qualitative research, was used. The study was carried out with 34 seventh-grade secondary school students. Data were collected and analyzed descriptively through observation, interview and various formative assessment tools used in the process. At the end of the three-week implementation period, substantial improvement was observed in the achievements of low and medium-achieving students, and the overall success of the students changed positively at the end of the process.

1. INTRODUCTION

The individuals who can manage and understand mathematics have the opportunity to use this knowledge in their business and daily lives in the future (Amankonah, 2013). Due to the unique spiral structure of mathematics, success in one achievement contributes positively to the achievement of the other achievement as well. Thus, it is inevitable that success brings success.

The idea of improving the mathematics achievement of students is even more important within the scope of the mathematics mobilization that has recently been initiated by the Ministry of National Education. Evaluation is the most important element in identifying the knowledge and skills of students that are effective in achieving these achievements. Even though evaluation has been perceived as an action taken at the end of the learning process for years, it has been realized that evaluation should be done throughout the learning process, and it has begun to be regarded as the most important element of learning. Even if all students are at the same level at the beginning of their education, students will be at different levels of cognition due to

*CONTACT: Hilal Özcan ✉ hilal-usta@hotmail.com Ministry of Education, Osmaniye, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

individual differences in the process, so evaluation turns out to be the most important element in effective teaching (Kültür, 2021).

Considering the necessity of process evaluation as well as result evaluation in the Turkish education system, the use of tools and methods called "alternative" and then "complementary" has been included in addition to traditional measurement and evaluation (MEB, 2020). In many studies conducted in our country on how competently the teachers use the alternative assessment tools, it is clear that there were many negative reasons such as the lack of financial means of schools and students, the attitude of parents, and the exam-oriented curriculum. It is obvious that since the curriculum was exam-oriented, written and oral exams were emphasized in terms of compliance with the system, which are the traditional assessments (Kâhtalı & Çelik, 2020) and that since central exams were result-oriented assessment, teachers had difficulty in applying the alternative assessment and evaluation approaches (Arseven, 2013; Karakuş & Yeşilpınar, 2013). Even though they were known by the teachers, they were not used. Therefore, it was emphasized in those studies that what formative assessment practices were like and how these activities could be applied should be demonstrated to the teachers through professional development studies (Arseven, 2013; Karakuş & Yeşilpınar; 2013; Önel et al., 2020).

Evaluation standards for school mathematics were introduced emphasizing that the assessment published in international standards was important not only what students knew or did not know, but also how they reasoned and what thought processes they adopted (Van de Walle, Karp & Williams, 2019, p.77). Four specific objectives were set, aimed at monitoring student progress, making decisions for instruction, assessing student achievement, and evaluating programs according to certain assessment standards. As far as the evaluations established for these purposes are concerned in line with their purpose, they are divided into three; diagnostic and placement purposes, summative purposes and formative purposes (MEB, 2020, p.5). Formative assessment, which is the subject of this study, is defined as the assessment carried out at every stage of the teaching process in order to support teaching, from seating arrangement to group work, supplementary worksheets, and activities throughout the teaching process (MEB, 2020).

There are two actions on the basis of the evaluation made for formative purposes; the first is the learner's understanding of the gap between the desired goal and his/her own purpose, and the second is the learner's action to close this gap in order to reach the desired goal (Black & Wiliam, 1998). Successful learning occurs when these two actions, which are the basis of formative assessment, are implemented. In many European countries, formative assessment is seen as an important strategy to acquire quality in education and guidebooks have been developed in order to guide teachers (Ozan, 2017).

In our country, Turkey, the "Mathematics Teacher Guide Booklet" prepared by UNICEF and the General Directorate of Assessment, Evaluation and Examination Services in 2020 focuses on formative assessment activities and serves as a guide for teachers in terms of application situations. Based on this guideline, the importance of demonstrating and disseminating the applicability of formative assessment has emerged in order to improve success in mathematics teaching.

In many international studies, it was reported that the formative assessment practices shaped the teaching practices by increasing student performance, that positive reflections were observed on students when teachers designed formative assessment practices effectively, that they were used to identify the gaps in learning, that the interactive formative assessments and feedback mechanisms contributed positively to students' understanding of the lesson, completing tasks, communication, interest in mathematics and social relations, and that the formative assessment approach increased students' success and metacognitive awareness compared to the traditional approach (Agwagah & Ezieke, 2023; Fatima, 2022; Hiloma & Briones, 2022; Martin et al., 2022; Miller, 2019; Wafubwa & Csikos, 2022).

In the national sense, even though there are a small number of studies on formative assessment practices, some studies on pre-service teachers and other disciplines have been encountered. Nevertheless, no other study was available, apart from Tekin's (2010) study, which was conducted for secondary school students on formative assessment-based teaching in mathematics teaching. This shows the deficiency in the number of studies in the relevant literature.

On the other hand, when we look at the national and international exams, it is clear that the course students are the most unsuccessful is mathematics. According to the High School Entrance Examination (LGS) 2022 report, mathematics was the lowest course with an average of 4.74 in the mathematics subtest compared to other courses (MEB, 2022a) and similarly, according to the 2018 PISA report, Turkey was in the 42nd place among 79 countries in the ranking (MEB, 2022b), despite an increase in mathematics achievement compared to previous years, and this explains that scientific studies should be carried out in order to improve the success of mathematics on a national and international scale.

In the mathematics course curriculum, teaching circle and circular region is included at all grade levels, starting with recognizing the circle in the 1st grade and ending with the achievements of the circle and circular region sub-learning area in the 7th grade (MEB, 2018). Although the acquisitions of calculating the circumference of the circle and circular region are an important part of the curriculum, many students do not have sufficient level of the concepts of perimeter and area measurement (Aksu, 2019; Görgüt, 2020). It is thought that the lack of conceptual learning about circles and circular regions may affect the geometric knowledge acquired later, and it is predicted that this problem, which will be experienced especially at the secondary school level, will be difficult to correct in the following years (Aksu, 2019). In his study on secondary school students' misconceptions, Kara (2021) stated that some students confused the definitions of circle and circular region and used them interchangeably, had misconceptions about the relationship between diameter and radius, and some students had misconceptions about determining the location of the center of the circle. In the study conducted by Evirgen and İkikardeş (2019), in which seventh-grade students received student opinions about the subjects they had difficulty with, it was observed that the students had difficulty with the circle and angles in the circle. Since the mathematics course has a spiral structure, students' learning deficiencies are negatively reflected in another subject. Therefore, studies should be carried out to ensure that students' learning is permanent. It is predicted that with formative assessment-based teaching, students will learn the concepts of circle and circular region more effectively and permanently.

Considering this low accomplishment in mathematics lessons and the mathematics mobilization started in our country, it is undeniable that different methods should be used to improve success in mathematics teaching. The general aim of this study is to improve students' mathematics achievement through assessment tools carried out in the context of formative assessment. Therefore, an exemplary application for teaching of the subjects of circle and circular region based on formative assessment was implemented in this study.

2. METHOD

In the study, action research was utilized as a qualitative research design. Action research is an approach that aims to analyze the systematic data collection in an attempt to raise questions about the implementation process or to understand and solve a problem that has already emerged, carried out by a practitioner working in a school, such as a manager, teacher, education specialist, directly or with a researcher (Yıldırım & Şimşek, 2021, p.319). In this study, action research was used as a method for the solution of mathematics failure in students by investigating the improvement of students' mathematics achievement with formative assessment-based teaching practice. In formative assessment-based teaching, the gap between the feedback received from the students and the teaching objectives and the knowledge acquired by the student is determined, new plans are made and developed on how to eliminate this gap, and the use of action research pattern in research to find solutions to the problems is in line with the nature of formative assessment in this context.

Since the researcher was also the mathematics teacher of the class in which he conducted the research, he took on the role of teacher researcher in this process. The researcher implemented the entire process himself. All progress was made by reaching consensus with the evaluation team before and during the implementation. At the same time, the field expert in the evaluation team is the researcher's thesis advisor. Process management was ensured by constantly obtaining expert opinion, taking into account scientific ethical principles. The formative evaluation tools used were revised in line with the opinions of the evaluation team. Each lesson was video recorded by the researcher. In addition, audio recordings were made of one-on-one interviews with students. The researcher personally conducted interviews and observations throughout the entire process. Observations during each lesson day and what was done in the lesson were noted by the researcher in the researcher diary. The data obtained throughout the research process was analyzed and interpreted by the researcher, progressed in line with the opinions of the field expert and evaluation team, and was reported by the researcher.

2.1. Study Group

Appropriate sample was used in the study. Appropriate sampling is defined as the collection of data from a sample that the researcher can easily access (Büyüköztürk et al., 2020, p.95). Therefore, the study was implemented with 36 seventh grade secondary school students at a state secondary school where the researcher taught. Two students who could not fully participate in the three-week application process due to health problems were excluded from the study. The remaining 34 students, 17 female and 17 male students participated fully in the implementation process. Analyzes of these 34 students were implemented.

It was identified from the readiness worksheets that all the students did not study and did not know the sixth-grade achievements about circle and circular region; therefore, the action plans were made based on this deficiency. When the students' mathematics achievement levels are examined, it is understood that their learning losses were high because they received distance education due to the pandemic the previous year. According to one-on-one interviews with students and school administration records, it was determined that only six to seven of the students regularly attended online classes. It was observed that the students with low success in mathematics courses also had low levels of readiness for the subject of circles and circular region, and when describing the students in the study. In this case, students whose mathematics grade point averages at the end of the first semester of the 2021-2022 academic year are between 0 and 54 are determined as low achievers, students with scores between 55 and 84 are determined as medium achievers, and students with scores between 85 and 100 are determined as high achievers (Table 1).

Table 1. *Mathematics achievement status of the students.*

Mathematics Achievement Status	Students
Low-achieving	S2, S3, S4, S5, S8, S9, S10, S13, S15, S16, S17, S20, S21, S24, S25, S26, S27, S28, S29, S30, S31, S33, S34
Medium-achieving	S1, S6, S7, S11, S12, S14, S18, S19, S32
High-achieving	S22, S23

As stated in Table 1, all students were coded as S1, S2, S3 …S34 based on the class attendance list, and the researcher was represented by using the letter A and the findings were reflected using these representations.

2.2. Data Collection Tool

In this study, in the three-week mathematics teaching implemented by using formative assessment tools, the success of the students regarding the subject of the circle and the circular region was aimed, and eventually, their success was attempted to be improved and their deficiencies, if any, were identified in the process. Therefore, observation and interview techniques were used throughout the process. Furthermore, formative assessment tools were used as data collection tools throughout the study process. Formative assessment tools (worksheet, rubric, performance-based assessment, checklist, quick techniques that can be used for formative assessment, observation form, self, peer and group assessment, concept maps, product file, dynamic software, digital media) were used by specifying the appropriate ones.

2.3. Data Analysis

The data obtained qualitatively were analyzed descriptively. With the formative assessment tools used in this study, the themes that contributed to the success of the students regarding the subject of circle and circular region were established. "Recognizing the center, radius and diameter by drawing a circle, Discovering that the ratio of the length of a circle to its diameter is a constant value, Calculating the length of a circle given its diameter or radius, Determining the relationship between the central angle, the arcs it sees and angle measurements" in the sixth and seventh-grade circle and circle teaching, The achievements of "Calculating the length of the circle and circle segment, Calculating the area of the circle and circle segment" were determined as the theme. The data obtained from the observations and interviews were coded, classified and presented under the relevant themes. Moreover, in line with the data obtained from the interviews, the mutual dialogues between the researcher and the students were represented with direct quotations.

2.4. Validity-Reliability

In terms of internal validity, the researcher is expected to be consistent both in the data collection processes and, in the analysis, as well as the interpretation of the data, and explain how this consistency has been achieved. While the researcher constantly questions herself and her research processes with a critical eye and checks whether the findings and results, he/she has obtained reflect the truth, making clear and understandable explanations to satisfy the reader provides internal validity (Yıldırım & Şimşek, 2021). In order to ensure the internal validity, various formative assessment tools were used in this study and the data collected from the students in the use of these tools were explicated in a clear manner.

Yıldırım and Şimşek (2021) reported that if the results of a study could be generalized to similar environments and situations, the research had external validity. Therefore, the fact that the study was conducted in a state secondary school, in which students with low, medium or high achievement in mathematics, was in a natural classroom environment for the whole class without making student selection demonstrated that the study had external validity. Even though the results obtained from the study cannot be generalized to all secondary school students, the experiences and observations obtained during the research process offer an exemplary application to formative assessment-based mathematics teaching.

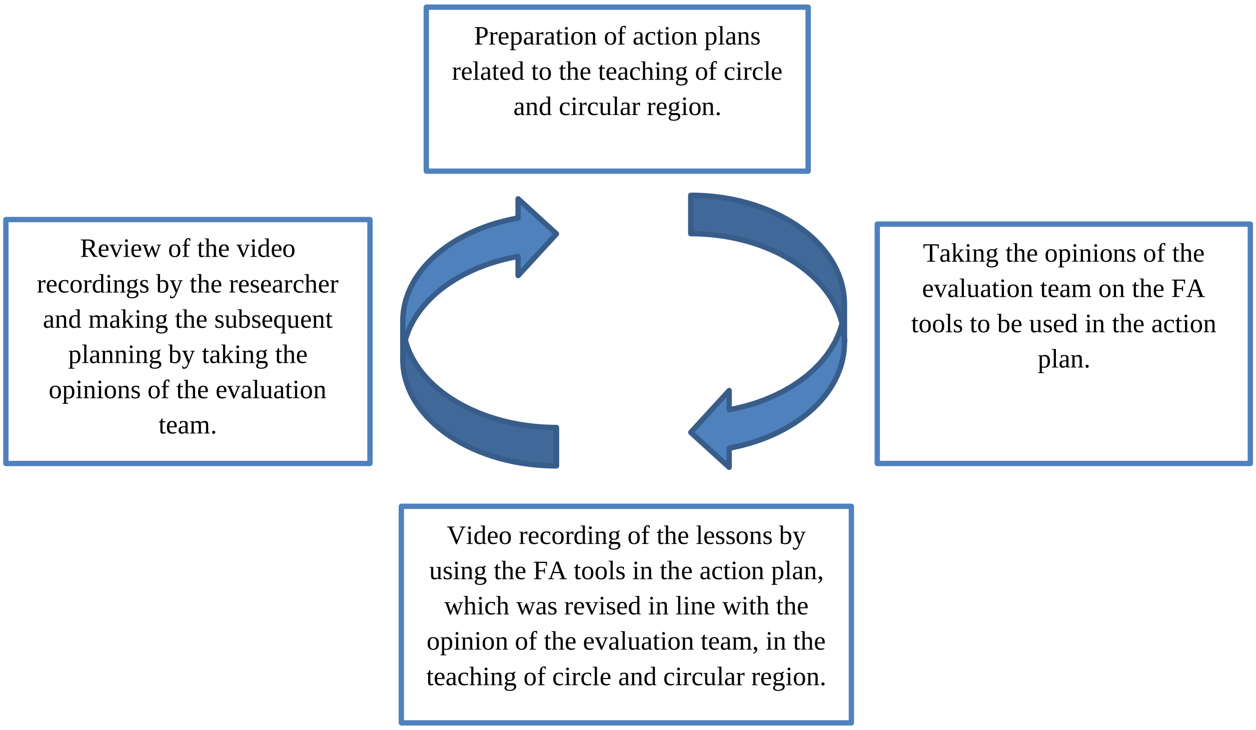
In order to ensure internal reliability in the study, firstly, one-to-one interviews with the students, the researcher's observations, the data obtained from the formative assessment tools used were directly analyzed and presented descriptively. Furthermore, before the research process started, an evaluation team consisting of two mathematics teachers teaching at the researcher's own school, a mathematics teacher teaching in a different state secondary school in the same province, and a field expert was formed. During the preparation of each action plan, the views of the evaluation team were obtained, and the necessary arrangements were made, and the plans were implemented. Moreover, the whole process was video-recorded, and revisions were made in line with the views of the field experts. Additionally, the data analyzes were compared with the field experts by making the analyzes. Thus, the obtained results were confirmed.

By reporting the stages, she followed in the study in detail and clearly, the researcher have shown that the results depend on the data she has collected and that her own assumptions or prejudices have not affected by the results (Yıldırım & Şimşek, 2021). In this study, in order to ensure external reliability, subjective judgments were avoided, and the contents obtained from the documents obtained from the students, the images obtained from the in-class video recordings, and the one-to-one interviews with the students were all presented as they were. What were done in data collection, processing, analysis, interpretation and reaching the results were all clearly explicated. The raw data of the study were saved to be examined by others and the whole application process was video recorded.

2.5. Action Research Process

The action research process consists of five basic steps: identifying the problem situation, planning data collection, collecting and analyzing data, preparing action plans in line with the findings, sharing and reporting the action plan and results by examining the relevant literature (Johnson, 2019). The action plans made during the study process were examined by an evaluation team of four people, including an expert in the field, two mathematics teachers teaching in the researcher's own school, and a mathematics teacher teaching in a different public secondary school in the same province. The nature of formative assessment (FA) is that it is a cyclical plan that is tailored to the student and is constantly revised according to the level of the student in the classroom. In this sense, it shows parallelism with action research. Therefore, action plans were constantly revised in line with the suggestions of the evaluation team according to the classroom climate, and action research process steps were implemented (Figure 1).

Figure 1. *Action research process.*



Preparation of action plans related to the teaching of circle and circular region.

Taking the opinions of the evaluation team on the FA tools to be used in the action plan.

Video recording of the lessons by using the FA tools in the action plan, which was revised in line with the opinion of the evaluation team, in the teaching of circle and circular region.

Review of the video recordings by the researcher and making the subsequent planning by taking the opinions of the evaluation team.

The action plans applied while following the action research process are as follows:

1. As an action plan, it was planned to apply a readiness test in order to measure the existing knowledge of the students about the subject of circle and the circular region. Students were given worksheets consisting of four questions and were asked to complete them according to the necessary instructions. In the first question; they were asked to draw a circle in the given unit on squared paper using compasses and ruler and show its radius and diameter on the figure. In the second question; they were asked to find the circumference lengths of circles given radius lengths and pi values, using relations. In the third question; they were asked to explain the difference between a circle and a circle. In the fourth question; they were asked to write the measures of a right angle (90^0), supplementary angle (180^0), and full angle (360^0). Student worksheets were evaluated and presented in Table 3. As a result of this test, it was obvious that the students were not at a sufficient level regarding the subject of circle and the circular region, more than half of the class could not attend the classes due to the pandemic, and they could not acquire the sixth-grade achievements.

2. As an action plan, after it was established that the students could not acquire the achievements of the sixth grade, one week (5 lesson hours) of the present study was allocated to the achievements of the sixth grade, with the suggestion of the teachers of the group and the expert opinion. Prepared lesson plans were reviewed by the evaluation team and required arrangements were made. In the first week, students were offered the learning achievements of recognizing the center, radius, and diameter of a circle by drawing a circle, discovering that the ratio of a circle's length to its diameter was a constant value, and calculating the length of a circle given its diameter or radius. Various formative assessment tools were used while offering these gains (See Table 2). Students were constantly evaluated with these tools and their weaknesses were observed and action plans were implemented for the achievements to be offered.

3. As an action plan, it was decided to move on to the seventh-grade achievements, as it was observed that the students acquired the achievements offered in the previous week with the formative assessment tools applied. With the views of the evaluation team, lesson plans were prepared to identify the relationship between the central angle, its arcs and angle measures, and calculate the length of the circle and the circle segment. The action plans were revised by continuously evaluating the students with formative assessment tools.

4. As an action plan, it was observed that the students achieved the previously given objectives, and a lesson plan was prepared together with the views of the evaluation team for the achievement of calculating the area of the circle and circle slice, one of the seventh-grade achievements. This was the last week, and the plans were revised by constantly evaluating the students using the necessary formative assessment tools.

Table 2. *FA tools used based on the learning outcomes during the implementation process.*

Implementation Process	Learning outcomes	FA tools used
1st week	• Recognizes its center, radius and diameter by drawing a circle. • Recognizes that the ratio of a circle's length to its diameter is a constant value by measuring it. • Solves problems that require calculating the length of a circle given the diameter or radius.	- Readiness worksheet - Worksheets - Group work and group work checklist -Activities: (Discovery of Pi-Poster making-Geometry board) and activity (Ferris wheel) - Homeworks - Quick evaluation (Plickers) - Performance based evaluation (Wordwall, EBA) - Information card - Quick techniques (Thumbs down-thumbs up, Red-green card) - Peer evaluation - E – portfolio
2nd Week	• Recognizes the relationships between the central angles, the arcs it faces and the angle measures in the circle. • Calculates the length of the circle and the circle segment.	- Self assessment - Team work - Homework (worksheet, textbook) - Quick evaluation (Plickers) - Quick techniques (Thumbs up-thumbs up, Red-green card, Question box) - Performance-based assessments and assignments (EBA, Wordwall) - Peer evaluation - Activity (Central angle and arc measure, Skill-based question study) - Separate homework assignments for certain people who are weak - E - portfolio
3rd week	• Calculates the area of the circle and circle segment.	- Performance-based assessments and assignments (Wordwall, EBA) - Activity (EBA, Skill-based question study-pizza slices) - Worksheets - Homework (Textbook, worksheet) - E – portfolio - Problem solving analytical rubric - Team work - Output card - Quick techniques (Thumbs up-thumbs up, Red-green card, Question box)

* FA: Formative assessment

In line with the prepared action plans, the study was conducted with a three-week implementation period, and the data obtained are presented in the findings section.

3. RESULTS

While the results regarding the improvement in the achievements of the students were presented, the readiness of the students was identified initially and the insufficiencies in the achievements of the sixth grade in the teaching of the subjects of circle and circular region were revealed. Therefore, the results of the circle and circular region teaching based on the formative assessment applied for the insufficiency of the students and the student achievement development towards the target that the students were expected to gain were presented. After the results related to the achievement development of the sixth-grade achievements, the findings of the student success improvements in the seventh-grade achievements were presented.

3.1. Results Related to Student Readiness and Students' Achievement Development Regarding Sixth Grade Achievements

In order to measure the readiness of the students, before the application process started, the students were given worksheets consisting of four questions and they were asked to answer it according to the required instructions. In the first question, they were asked to draw a circle in the given unit on the squared paper using a compass and ruler and show its radius and diameter on the circle. In the second question, they were asked, using the relation, to find the perimeters of the circles, whose radius and pi values were given. In the third question, they were asked to explain the difference between a circle and a circular region. In the fourth question, they were asked to write the measure of right angle (90^0), supplementary angle (180^0) and full angle (360^0). Student worksheets were evaluated and presented in Table 3 can find a sample table and a figure presented here.

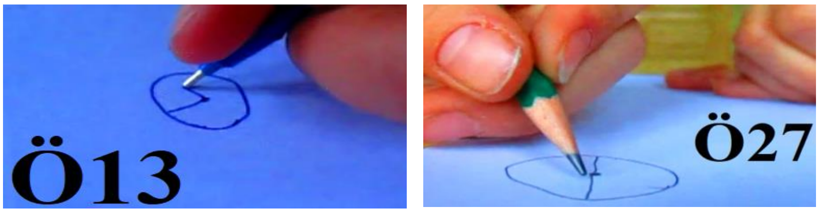
Table 3. *Evaluation of students' responses*.

Questions	Learning outcomes	Students who failed to acquire any learning outcomes	f
1.Question	Able to show center, radius, diameter on a circle.	S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11,S13,S15,S16,S17,S18,S20,S21,S24,S25,S26,S27,S28,S29,S30,S31,S32,S33, S34	29
2.Question	Able to calculate the circumference of a circle.	S2,S3,S4,S5,S6,S8,S9,S10,S13,S15,S16,S17,S20,S21,S24,S25,S26, S27, S28,S29,S30,S31,S32,S33,S34	25
3.Question	Able to explain the difference between a circle and a circular region.	S4,S5,S8,S9,S15,S18, S20, S26, S27,S28,S30, S31,S33	13
4.Question	Able to write perpendicular, supplementary and full angle degrees.	S9,S10,S13,S20,S26,S31	6

As is clear in Table 3, more than half of the class failed to draw the circle given in a certain unit and failed to show its center, radius and diameter. 25 students in the class could not find the circumference of the circle given the radius. This fact proved that the students failed fully learn the concept of the circle and could not make sense of the pi number we used to calculate the circumference of a circle. Looking at the students' answers to the third question, it was clear that 13 students failed to explain the difference between a circle and a circular region. It was observed that 21 people who were able to explain made a comment from memory as "*the circle is empty; the circle is full*". *Finally, 6 students could not write the concepts of right angle, supplementary angle, and full angle, and 28 students wrote "right angle 90 degrees, supplementary angle 180 degrees, full angle 360 degrees*".

Considering the readiness of the students, the first week of the application process was given priority to the achievements of the sixth grade and the achievements of "Recognizing the center, radius and diameter by drawing a circle, discovering that the ratio of a circle's length to its diameter is a constant value, Calculating the length of a circle given its diameter or radius" were attained. It was observed that the students with low mathematics achievement acquired the achievements of showing the center, radius and diameter by drawing a circle in the process (Figure 2).

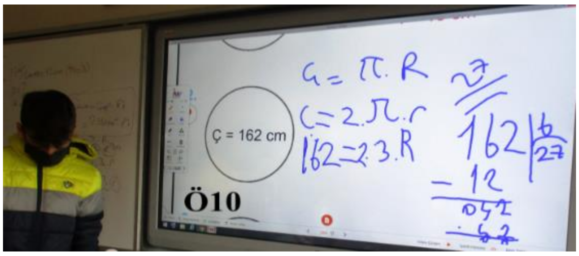
Figure 2. *Drawings of the student coded S13 and S27.*



The fact that the students coded S13 and S27, with a very low success in general, and never learnt the subject before were asked by the researcher to draw the circle and show its radius and diameter, one the acquisitions of drawing a circle and showing its radius and diameter, were able to draw a circle and show its radius and diameter, was an important indicator of success for these students.

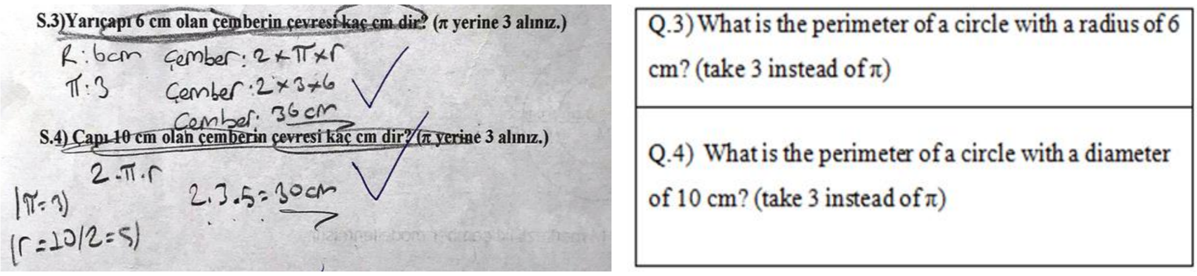
Considering that more than half of the class had never acquired the achievement of *"calculating the length of a circle with a given diameter or radius"* since the students had problems participating in the lessons during the period when the lessons were conducted with distance education due to the pandemic, it was observed that the fact that almost all of the class acquired the achievement with the formative assessment-based teaching applied in the process illustrated the improvement. In the formative evaluation process, being able to solve the questions and improving him/herself in line with the feedback given by the teacher led to an improvement in the performance of the students. Notably, the students with normally low in-class performance increased their success as they experienced the feeling of being able to correctly answer the questions and success (Figure 3).

Figure 3. *Class participation of the students with low in-class performance.*



As is clear in Figure 3, when the readiness level of the student coded S10 with low in-class performance was measured, it is obvious that he could not calculate the circumference of a circle whose radius or diameter was given (See Table 2). With the assessment tools, the student with the code S10 acquired the achievement and demonstrated this in the classroom. By making explanations on the board by the researcher teacher both individually and for the whole class in the classroom, the students were asked to perform individually, students' deficiencies were detected on the digital applications and their e-portfolios, and they were able to make up for their deficiencies in the achievements thanks to the feedback provided. For instance, it was observed that the student coded S28, whose success in mathematics was low, lacked in the acquisitions of being able to show the center, radius, diameter on the circle in readiness, to calculate the circumference of the circle, and to explain the difference between the circle and the circular region (See Table 2). Considering this deficiency, with the help of group work, worksheets, activities, homework, Plickers, Wordwall, EBA platforms, rapid techniques, peer assessments and various formative assessment tools applied in the form of flashcards, the student coded S28, like every other student, was also able to prove his improvement on paper, which was an important indicator of success (Figure 4).

Figure 4. *Turkish and English versions of the worksheet of student coded S28.*



S.3)Yarıçapı 6 cm olan çemberin çevresi kaç cm dir? (π yerine 3 alınız.)

R:6cm Çember: 2×π×r

π:3 Çember: 2×3×6

Çember: 36 cm

S.4) Çapı 10 cm olan çemberin çevresi kaç cm dir? (π yerine 3 alınız.)

2.π.r

(π=3) 2.3.5=30cm

(r=10/2=5)

Q.3) What is the perimeter of a circle with a radius of 6 cm? (take 3 instead of π)

Q.4) What is the perimeter of a circle with a diameter of 10 cm? (take 3 instead of π)

As seen in Figure 4, with the formative evaluation activities carried out during the process, the student with the code S28, using the relation, was able to calculate the circumference of the circle whose radius was given in the first question; in the question below, he realized that he had to use the radius length of the circle, whose diameter was given in the relation, and was able to calculate the radius length and calculate the length of the circumference of the circle. The fact that the student coded S28 could make sense of the concepts of radius and diameter and use them in the calculation of the circumference of the circle, knowing how to write them on the worksheet, proved the success of the student.

At the end of the one-week period, the worksheet consisting of 10 open-ended questions containing the achievements of the sixth-grade was given to the students and the results are presented in Table 4.

Table 4. *Mid-term evaluation achievement distributions.*

Mid-term Evaluation	f (Frequency)	(Percentage) %
Those with 6 or more correct answers	25	76%
Those with 5 or fewer correct answers	8	24%

When the answers given to the worksheet, consisting of open-ended questions including the achievements of calculating the radius and diameter length, understanding the meaning of the number pi, and calculating the circumference of the circle, were examined, it turned out that the class had a success rate of 76%. Individual assignments were prepared for the students with deficiencies, a separate group was created on the EBA platform, video narrations and study questions were sent for the missing achievements, and these students were followed up and the seventh-grade achievement was initiated by eliminating the deficiencies of these students.

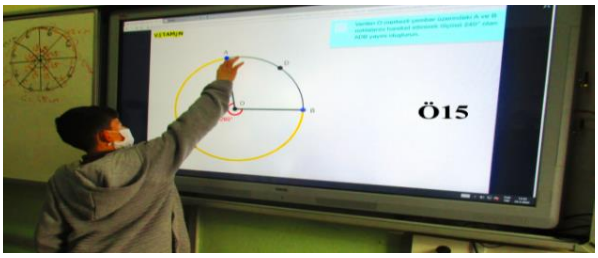
3.2. Results Related to the Success Regarding the Seventh-Grade Achievements

Results related to the success regarding the seventh-grade achievements were presented under the sub-headings of identifying the relationship between the central angle, the arcs it faced and angle measures, calculating the length of the circle and the circle segment, and identifying the area of the circle and slice of the circle.

3.2.1. *Determining the relationship between the central angle, the arcs it faces, and the measures of the angle*

Regarding the seventh-grade achievements, the students clearly got accustomed to the application process, and they managed their own learning by taking an active role in the course content enriched with various formative assessment tools. Considering the state of the classroom before the implementation process, few students participated in the lesson, while the whole of the class actively participated in the formative assessment activities throughout the process. For instance, the student coded S15 was an average success student, and while his in-class participation was weak, he fully participated in the activities held during the process (Figure 5).

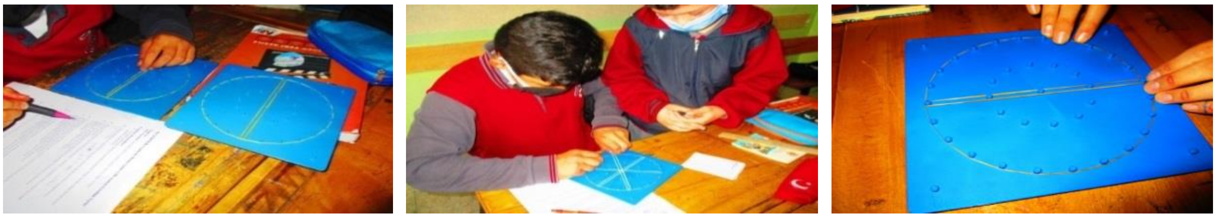
Figure 5. *Digital activity participation.*



As is clear in Figure 5, the student came to the board and observed the relationship between the central angle and the arc measure dynamically on the smart board in the digital application. The student's performance development was evaluated as formative by asking different questions by the researcher teacher. With the immediate feedback given by the researcher after the answers received from the student, the student coded S15 improved his own success by making sense of the relationship between the central angle and the arc he saw.

In addition to the digital materials, concrete materials such as geometry boards were used to identify the relationship between the central angle in the circle and the arc measure seen in the circle, and various in-class group activities were used to enable the students to learn more effectively (Figure 6).

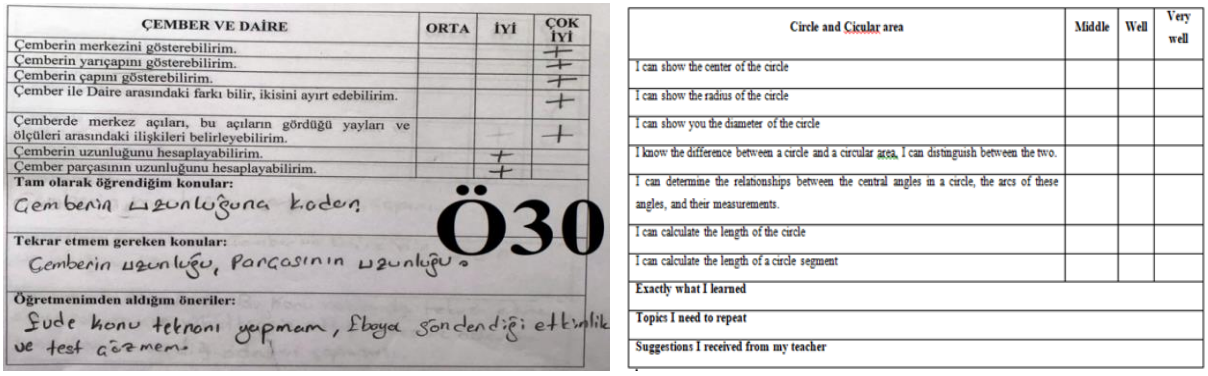
Figure 6. *Images related to the activity.*



In the formative evaluation performed using the geometry board, the relationship between the angle and the arc was observed through group studies, as is observed in Figure 6. By forming groups of two, following the instructions on the given worksheet, the students were allowed to discover the central angle and the arc measure seen by the central angle on the geometry board. After the students, who followed the instructions in the worksheet by proceeding in a semi-circle, quarter-circle, answered the questions on the worksheet by discussing it with their friends, the researcher teacher proceeded in the form of a question and answer, and it was established where the students had deficiencies. The students, whose deficiencies were observed, were given immediate feedback by the researcher teacher, and the students' deficiencies were eliminated, and they were enabled to acquire the achievement.

In some of the group studies carried out during the process, peer assessment and self-assessment were used as formative assessment tools. The peer assessment and self-assessment used were influential in the teacher's planning of the next lesson and having information about the students (Figure 7).

Figure 7. *Turkish and English versions of the self-evaluation form of the student coded S30.*



ÇEMBER VE DAİRE	ORTA	İYİ	ÇOK İYİ
Çemberin merkezini gösterebilirim.			+
Çemberin yarıçapını gösterebilirim.			+
Çemberin çapını gösterebilirim.			+
Çember ile Daire arasındaki farkı bilir, ikisini ayırt edebilirim.			+
Çemberde merkez açıları, bu açıların gördüğü yayları ve ölçüleri arasındaki ilişkileri belirleyebilirim.			+
Çemberin uzunluğunu hesaplayabilirim.		+	
Çember parçasının uzunluğunu hesaplayabilirim.		+	

Tam olarak öğrendiğim konular:
Cemberin uzunluğuna kadar

Tekrar etmem gereken konular:
Cemberin uzunluğu, Parcasının uzunluğu.

Öğretmenimden aldığım öneriler:
Şude konu tekrarı yapmam, Ebaya göndendiği etkinlik ve test çözmem.

Ö30

Circle and Cicular area	Middle	Well	Very well
I can show the center of the circle			
I can show the radius of the circle			
I can show you the diameter of the circle			
I know the difference between a circle and a circular area. I can distinguish between the two.			
I can determine the relationships between the central angles in a circle, the arcs of these angles, and their measurements.			
I can calculate the length of the circle			
I can calculate the length of a circle segment			
Exactly what I learned			
Topics I need to repeat			
Suggestions I received from my teacher			

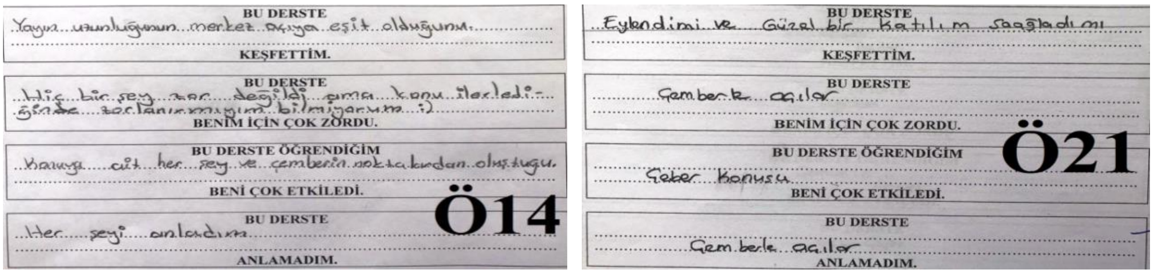
After the process of acquiring achievements was completed, a self-evaluation form was prepared for the students, as seen in Figure 7, and the students were asked to evaluate themselves. In the self-assessment of the student, coded S30, with poor mathematics achievement, he could show the center, radius and diameter of the circle very well, learn the difference between the circle and the circle, identify the relationship between the center angles in the circle and the arc measures seen by these angles, and calculate the length of the circle and the circle segment well. He also emphasized that he should study again the last subject of the circle and the length of the circle segment.

The student coded S30 saw his shortcomings in his self-evaluation and stated what he should do about his deficiency. Therefore, additional study questions were given to the student coded S30 and his efforts were followed up. In this way, the researcher teacher made up for the deficiencies of the students by identifying the subjects that the whole class had deficiencies in, planning the next lesson, giving homework for the students who were missing and following up. The self-assessments used because the class was crowded were an effective formative assessment tool in guiding the teacher by providing extensive information about the subjects covered in a short time about the students.

In addition to the self-evaluation form, individual interviews were held with the students who showed success in the lesson at the end of the lesson. In the individual interviews, the statement of the student coded S13, *"I used to be worse in this subject, but now I started to get better. I didn't solve many questions in the past, now I started to solve them, I started to like it a little bit, now it's been better now..."* was the indication that the student could make his own self-assessment and these self-assessments were related to the existing competence level of the student. Another student coded S16 said, the statements *"I don't know, I haven't been doing homework properly since I started 7th grade, that is, I didn't study enough for my lessons. Since these activities have been available, the photocopy worksheets and the things you downloaded from the internet were all good, math is the only I study willingly now, it will be better if it continues like this until the 8th grade"*, showed that the study performance of the student who had low success in the course and was not interested in the course changed for the better.

The formative assessment-based teaching implemented clearly contributed to the improvement in the student's success by positively changing the student's study performance. Another instance: the statement of the student coded S26, *"Actually I don't like to solve so many questions at home, especially in mathematics, but for some reason, I happen to a desire to solve questions at home about this circle,"* created the feeling of "I am able to" in the student since the main theme was to evaluate the student frequently and observe his deficiencies in the form of formative assessment-based teaching, taking steps appropriate to the level of the student. Furthermore, the researcher's observations during the process and the students' self-evaluations demonstrated the efficiency of formative assessment-based teaching in terms of planning the following lesson. Such evaluations also generate success. The statements of student coded S20 as *"I learned a lot thanks to friends"* and that of student coded S4 as *"Elif explained the subject to me, I understood it, Ma'am"* illustrated the contribution of peer learning to success in formative assessment. In addition to the self- and peer-assessments, the information cards given to each of the students (Figure 8) determined what the students knew and did not know about the achievement that was explained on that day' and additional activities such as in-class lectures, sharing lecture videos via the digital platform, and individual assignments were made available to the weak students.

Figure 8. *Information cards for students coded S14 and S21.*



BU DERSTE
Yayın uzunluğunun merkez açıya eşit olduğunu
KEŞFETTİM.

BU DERSTE
Hiç bir şey zor değildi ama konu ilerledi-ğinde zorlanırmıyım bilmiyorum :)
BENİM İÇİN ÇOK ZORDU.

BU DERSTE ÖĞRENDİĞİM
kavuşa ait her şey ve çemberin noktalardan oluştuğu.
BENİ ÇOK ETKİLEDİ.

BU DERSTE
Her şeyi anladım
ANLAMADIM.

Ö14

BU DERSTE
Eylendimi ve Güzel bir katılım sağladım
KEŞFETTİM.

BU DERSTE
Çemberde açılar
BENİM İÇİN ÇOK ZORDU.

BU DERSTE ÖĞRENDİĞİM
Çeber konusu
BENİ ÇOK ETKİLEDİ.

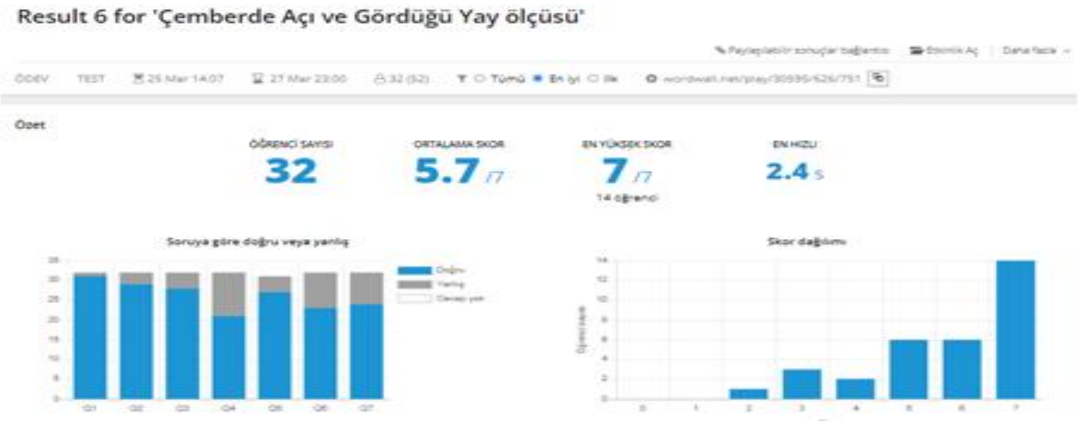
BU DERSTE
Çemberde açılar
ANLAMADIM.

Ö21

In the information card of the student coded S14, as seen in Figure 8, it is obvious that the length of the arc was equal to the central angle for that lesson, that he comprehended everything and that nothing was difficult for him. On the information card, the student coded S21 stated that he had fun in the lesson, participated, had difficulty regarding the subject of angles in the circle and did not understand it. By analyzing these two student profiles, the researcher teacher established what the students needed individually and planned the next lesson accordingly. While feedback was given to the student coded S14 that the concepts of arc length and arc measure were different, additional explanations were made for the student coded S21 and students with deficiencies like this student. In order to make up for the deficiencies, homework and digital activities were allocated and followed up, and the success development of the students was ensured.

In addition to concrete materials, digital platforms such as Geogebra, EBA (Educational Informatics Network), Wordwall were used to improve the success of students in a versatile manner. While some of the students fully participated in the practice activities sent via EBA during the process, some did not. As a result of the interviews with those who did not participate, negative situations were encountered such as the EBA platform having systemic problems, the fact that most of the students did not have a computer, so they had to log in on the smart phones and their phones did not have enough equipment to use the EBA platform. The statement of the student coded S17 as *"EBA is not available, I cannot access EBA, the smart phone cannot handle it"* and the statement of the student coded S16 as *"We have problems from EBA because we cannot access it; so, is better, do not send the practice activities on EBA, if you send 5 or 6 more activities on Wordwall, we will do them as well."* also represented the general level of the class as a whole. Therefore, in addition to the practice activities sent from EBA, Wordwall platform was also used as an opportunity for each student to have easy access. It was observed that student participation was higher because it was a game-based platform that students could easily access on their smart phones via the link. Assignments in the form of 10–15-minute quizzes were sent as links in a way that would not take the students' time and would attract their attention. The reports of the given assignments were reflected on the screen on the smart board and the students were given feedback collectively or individually (Figure 9). Giving feedback immediately was important in terms of thrusting the responsibility of homework of the students to and ensuring that the deficiencies were dealt with and corrected immediately. Moreover, using homework as a formative assessment tool in a crowded classroom also saved time.

Figure 9. *Report image of the Wordwall platform.*



Result 6 for 'Çemberde Açı ve Gördüğü Yay ölçüsü'

Özet

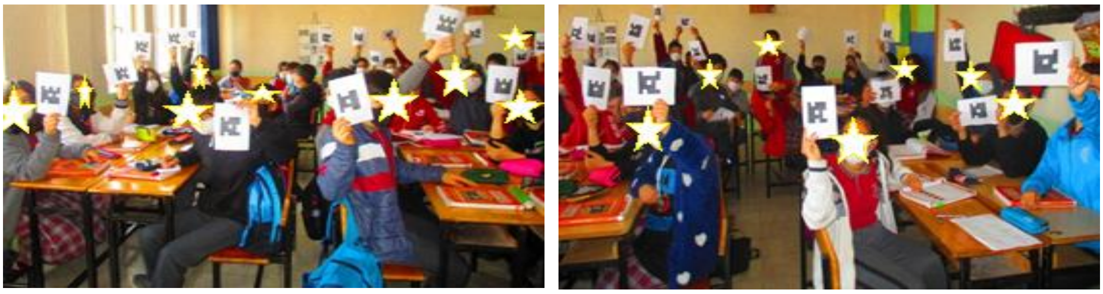
ÖĞRENCİ SAYISI	ORTALAMA SKOR	EN YÜKSEK SKOR	EN HIZLI
32	5.7 /7	7 /7 14 öğrenci	2.4 s

As is clear in Figure 9, all the students were able to use this application and the following day, practice activities were carried out on the deficiencies of the students in the classroom. Hence, the success improvement of the students, who were frequently evaluated with formative assessment tools and given feedback, gained momentum.

The statistics offered by this program were recorded as the e-portfolio of the students and the progress of the students was followed up through these reports. Apart from the applications made on paper, digital platforms increased the motivation of the students as well. Statement of the student coded S26 *"It is easier in digital because sometimes I cannot easily find what is available in the book and where it is, or I have a problem. If there is information somewhere, I rack my brain for hours trying to remember and find which page that information was on, it gets very confusing.",* in fact demonstrated that giving homework from digital applications about what he learnt in the lesson on that day was clearer and more efficient for the student to know what to do. Similarly, the progress in the success of the students was observed more clearly and quickly. The positive performance of the students who received feedback on similar questions in the following practice activity was an indication that their success increased.

Moreover, the plickers application, which gives immediate feedback to the students in the classroom, was used to measure the students' readiness for the previous acquisition at the introduction to the course or for a short review at the end of the course (Figure 10).

Figure 10. *Plickers application.*

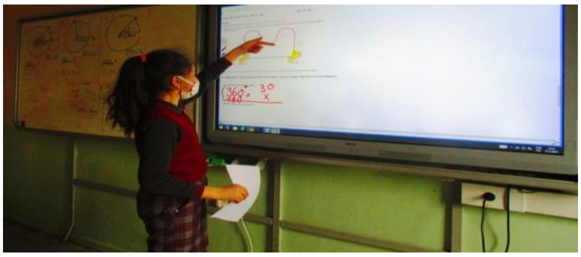


In this application, the whole class actively participated in the lesson, as each student lifted their own paper in order to answer the questions asked and saw his/her name on the board. In fact, the students who did not participate in the course at all (S8, S9, S17, and S25) felt more confident in this practice, which was one of the factors affecting the success of the students. As an indicator of their sense of security, the statement of the student coded S12 about the increase in the participation of the student coded S25, whose success in the course was very low, can be given as an example: *"I think it is a very useful activity, you don't criticize people; for instance, if you do it wrong, people laugh, and I get pissed about it. That's why I like that you do* *something like this. For instance, take a look at S25, because he doesn't study much, they laugh at him for doing it wrongly. But no one knows because you make a mistake. This caught my attention a lot."* Similarly, the statement of student coded S6 *"They laugh when someone makes a mistake, sir; but, it is not clear who makes a mistake here"* and the statement of student coded S7 as *"I think the QR code activity is very good, everyone lifts it up directly, there is no chaotic sound concentration"* explicate the positive aspects of this practice in terms of allowing the shy students in expressing themselves better in the lesson, being able to answer questions more easily, being practical and giving immediate feedback to the student. Furthermore, the fact that students did not need a phone in this application demonstrated that the plickers application was a useful tool as a formative assessment tool.

3.2.2. *Calculating the length of the circle and the circle segment*

The students who learned how to calculate the circumference of the circle by giving the radius and diameter length were provided with various formative assessment tools in order to calculate the length of the circle segment. For instance, another tool used in the process was peer review. It was also an indicator of success that there were students whose self-confidence increased with the explanation of their friends and who never wanted to go to the blackboard for an activity but learnt with the explication of their friends and wanted to do the activity voluntarily (Figure 11). For instance, the student coded S27, who did not go to the blackboard and wanted to go that moment, explicated herself as in the following:

Figure 11. *Student codes S27 with increased self-confidence.*



R: You don't usually go to the board, I had you come to the board last time, but you went and sat back without solving it. Well, now why did you go to the board on your own accord?
S27: I asked Meryem and she explained it to me. My self-confidence increased when my friends explained it to me, that's why I went to the board.

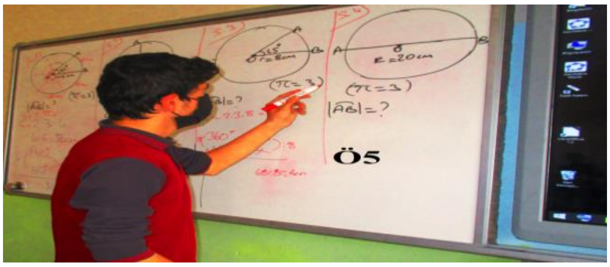
The observation of the student coded S7 as *"Hasan and Tugra never studied for their lessons, they simply ask each other when a group is formed, they learn because they explain one another what they do not understand"* illustrated that not only the researcher but also the students noticed each other's progress.

In some cases, on the other hand, the fact that a friend voluntarily helped a weak student extracurricularly and explained the subjects she could not understand was another indicator of peer learning (Figure 12).

Figure 12. *Extracurricular peer learning/teaching***.**

Here, the fact that the student coded S5 wanted to observe his own effort and wanted really to learn, previously as a student who did not willingly go to the blackboard at the normal time' but now more willing to make the effort showed the success of the student (Figure 13). Considering that the basis of formative assessment is to evaluate each student according to their own potential, this finding showed the positive change in the student.

Figure 13. *The student coded S5 is able to show that she has learned successfully***.**



As is clear in Figure 13, the student coded S5 calculated the length of the circle segment and was able to show it individually.

It is undeniable that the students' frequent exchange of knowledge and information with their groupmates in the group activities carried out during the process contributed to their success. So much so that, in the interviews with the students, the positive aspects of group work on themselves were indicated by the following statements of the students:

"Everyone can make up for each other's deficiencies in the activity, one helps the other when s/he cannot do it, and mostly these things are easier with cooperation." (S1)
"They couldn't do it, I helped them, and I showed them how to do it." (S6)
"I also like teaching to my friends; I reinforce what I teach as well. I think when the group is formed, everyone asks each other, and almost everyone can get on well with each other." (S7)
"Normally, I would cooperate with Kevser in such activities, I didn't communicate much. This time, I liked it when I did a nice project with all my friends. Thanks to this, I communicated more, I started to like my friends, and I always liked to communicate with them." (S11)
"Initially, I couldn't get along with some friends, afterwards, I started getting along better with them." (S17)
"I discuss questions with friends, they ask me as well and I help them." (S18)
"I learned to solve questions I didn't know from different perspectives. I also got different opinions of my friends and added them to my own, so I had an extra idea." (S23)
"For instance, we got along better." (S31)

Considering the comments of the students in general, it is possible to say that the contribution of peer learning to success was high. As a result of these positive feedbacks, various group activities as double, triple and hexadecimal groups were included throughout the implementation process, depending on the nature of the activities (Figure 14).

Figure 14. *Group work.*

As is clear in Figure 14, the students carried out their activities in groups of two and six according to the instructions in the worksheet given by the teacher. Students with a low level of success in the group showed improvement in their success by benefiting from the students with a high level of success. For instance, the dialogue with the student coded S12 is given below:

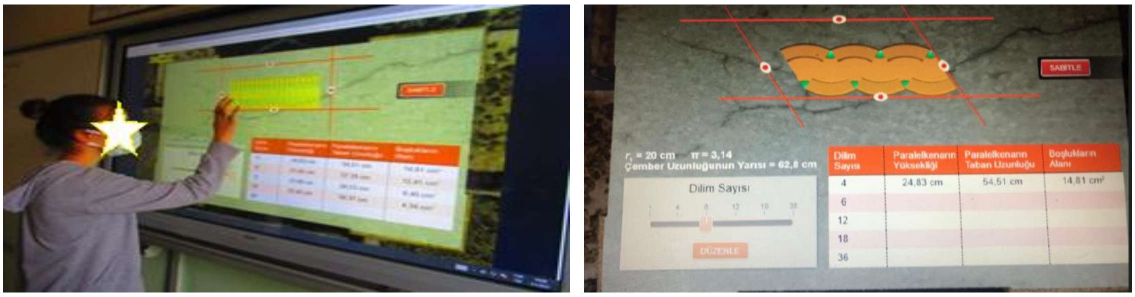
R: Do these two-person or six-person activities make any contribution to you?
S12: Yes, they do quite a lot. It's because there may be subjects that I do not know, I ask my friends, they teach me, this is how I learn.
R: Well, they make contributions to you, as far as I understand, you also make contributions to them.
S12: Yes, quite a lot
R: Well, is that a good thing?
S12: I think it's been great this way.
R: Actually, they criticized the group a lot regarding the seating arrangement,
S12: I think it's good in terms of sharing, the problem is that they talk a lot. Everything else is fine, when he asks about something he doesn't understand, they teach him one by one, if he doesn't understand, they ask someone else, s/he explains.

In addition to the positive aspects, the fact that there were disagreements and arguments in some groups according to the dynamics of the groups also demonstrated the negative side of this formative evaluation practice.

3.2.3. *Calculating the area of a circle and a circle segment*

After learning how to calculate the length of the circle segment, students started to learn how to calculate the area of the circle and the circle segment. The activity seen in Figure 15 in the Education Information Network (EBA) platform was applied to the students individually and the students were allowed to explore the area of the circle. During this exploration of the students, the students progressed with the questions asked by the researcher teacher, and feedback was given instantly according to the student's answers, allowing the students to discover the relation of the area of the circle.

Figure 15. *Individual in-class activities.*



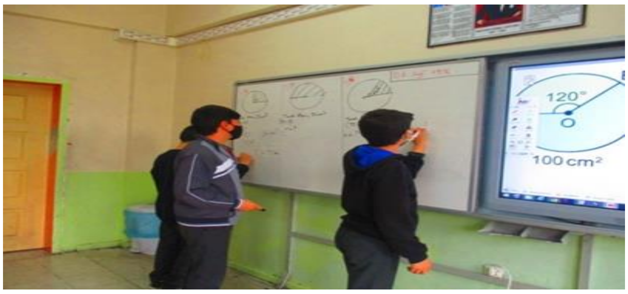
r_1 = 20 cm π = 3,14
Çember Uzunluğunun Yarısı = 62,8 cm

Dilim Sayısı	Paralelkenarın Yüksekliği	Paralelkenarın Taban Uzunluğu	Boşlukların Alanı
4	24,83 cm	54,51 cm	14,81 cm²
6			
12			
18			
36			

In the activity in Figure 15 within the EBA platform, it is clear that the student coded S26, who had low mathematics achievement and had low in-class motivation before the application process, improved her in-class motivation and participated in the activity. In this activity, the students discovered the transition from the area of the parallelogram to the area relation of the circle by increasing the number of slices.

During the in-class activities, the students received immediate feedback, saw where they had deficiencies, and managed the process accordingly. In some cases, feedback was given individually while the students were solving on the board in some cases (Figure 16).

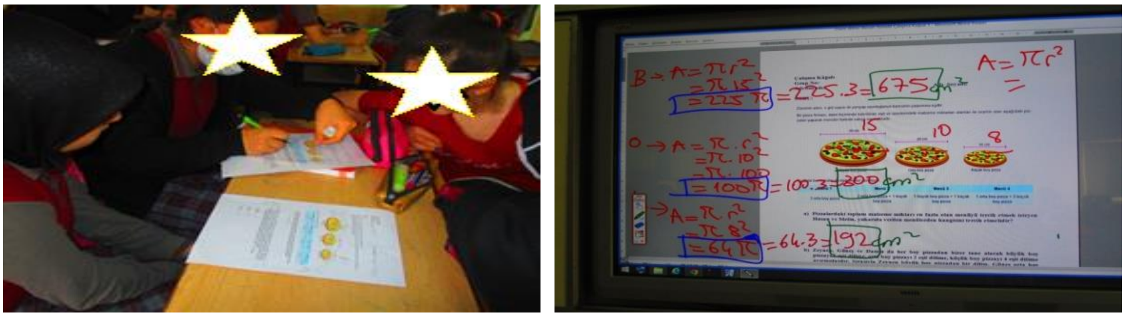
Figure 16. *Instant feedback to students.*



As is clear in Figure 16, the students who were identified to be deficient in in-class formative assessments were individually asked to come to the board and additional questions were asked to them. While the students were solving the questions about finding the area of the circle on the board, a formative evaluation in the form of question-answer was implemented, and the students were given feedback according to their answers, and the students were able to find the area of the circle and the circle slice.

In addition to the individual activities, in-class activities (Figure 17) performed by the students in pairs or trios were also used as a formative assessment tool in the process.

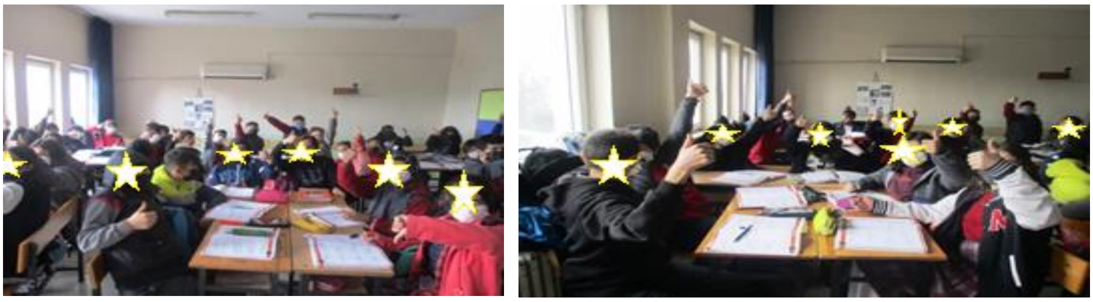
Figure 17. *In-class group work.*



In Figure 17, the skill-based question, which included finding the amounts of pizza and pizza slices associated with daily life, was asked to the students who formed groups of two by cascading. After the students completed their worksheets in the given time, the questions were discussed and resolved on the smart board. In the study, which was carried out with the participation of the whole class, the students answered the questions with peer learning, and the researcher teacher gave feedback while the students were answering the questions on the board or walking around the classroom. Giving feedback and enabling students to see where they made mistakes positively contributed to the improvement of their success.

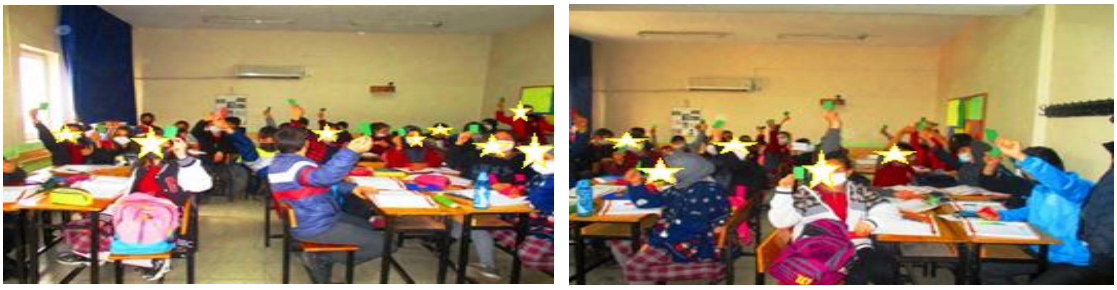
During the study process, the use of *thumbs up and thumbs* down (Figure 18), one of the formative assessment methods as instant feedback, was effective in terms of instantly recognizing whether students understood and in terms of working on it, and as a result, the students were always active in the lesson.

Figure 18. *Thumbs up thumbs down.*



If the students correctly understood the subject of circle and the segment of the circle slice at the moment of teaching, they put their thumbs up, if they didn't, they put their thumbs down. Even though the thumb was held sideways, it meant that they understood half, and those who had comprehension problems were helped to work on it again. The use of the red and green card used in the process was also one of the formative assessment tools used in the instant assessment (Figure 19).

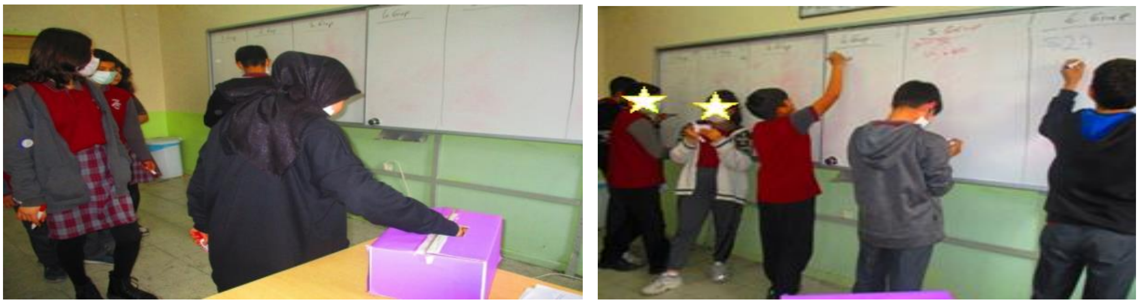
Figure 19. *Use of red green cards.*



As is clear in Figure 19, it was a formative assessment tool performed by holding up the red, green card, the green card if the subject was understood, the red card if not understood, and both if it was half understood. Furthermore, the use of these cards to give feedback to their friends who correctly solves the question on the board should also be used by the students to know the subject, which then leads to success.

Many formative assessment tools were utilized in the process, and even the formative assessment tools used were reformatted according to the students' competence and various external factors. For instance, students were asked to create a question box by throwing their own questions into a box, and the questions were drawn from the boxes and solved on the board in the time given in groups in the class (Figure 20).

Figure 20. *Question box formative assessment tool.*



The question box tool used in this activity was composed of questions about the circle and the area of the circle segment that the students prepared themselves. Initially, the question box activity, which was applied after the circle and arc length acquisition were taught to the students, was applied secondly after the acquisition related to the area of the circle was comprehended. In the second application, which consisted of questions about the area of the circle, when the students with low success levels did not want to volunteer to solve it in the first application, the instruction was changed, and it was suggested that students who could not solve the question could go to their own groups and learn and solve them on the board (Figure 21). Question box activity was formatted according to students and classroom climate.

Figure 21. *Students who couldn't solve getting help from friends.*

As is clear in Figure 21, all the students who learned with the help of their friends with the question box activity actively participated in the lesson and their success levels clearly improved.

4. DISCUSSION and CONCLUSION

As a result, in the present study, it was investigated how the formative assessment-based application process in the seventh grades in mathematics teaching regarding the subject of circle and the circular region improved the success of the students and an exemplary application for the formative assessment-based teaching in mathematics teaching was presented. Various formative assessment tools used in the study enabled the students to recognize their deficiencies on the subject at that moment and paved the way for making plans in an attempt to make up for the student's deficiencies. Therefore, the students both increased their motivation by tasting their own success instantly, and the enhancing interest in the lesson improved the success of the lesson. It turned out to be clear that the process evaluation brought about success as the students progressed with the activities suitable for their level and were frequently evaluated and their deficiencies were made up for. The results obtained from this study also overlapped with the findings of previous studies that formative assessments were an important factor that augmented the learning levels of secondary school students and had a positive effect on their success (Austin Hurd, 2015; Kline, 2013; Collins, 2012; Tekin 2010).

As far as the results of this study are concerned, it was revealed that not only the students with high mathematics achievement, but also low and medium successful students made progress in this process with the FA tools used in the teaching of the subject of circle and circular region. With various FA tools used in the teaching of the subject of circle and circular region based on formative assessment, clear contribution was made to the formation of a participatory classroom environment, to the progress in success despite a crowded classroom, to more active participation of students with low achievement levels and the non-participation in this process, to the increase in their learning motivation (Black & Wiliam 1998; Burns et al., 2010; Miesels et al., 2003), to the increase in the self-confidence of students with high success levels by sharing the knowledge with low-achieving students, to the students’ ability to recognize where they were weak and to manage the learning process themselves thanks to feedback, to the students' adaptation to the process at the end of two weeks and the demand for the lessons to be always like this, most importantly, to the students' feeling of being able to do and their participation and motivation to increase, and accordingly, to a decrease in their anxiety towards mathematics and to the achievement of success. This result overlaps with the result that formative assessment significantly increased students' academic success in many international studies (Kline, 2013; Burns et al., 2010; Foster & Poppers, 2009). The fact that the study significantly increased the success of students is in line with the result of Tekin's (2010) study that formative assessments in secondary school mathematics teaching had a positive effect on students' learning speeds and achievements.

Furthermore, the application of the study in a crowded classroom environment in a public secondary school demonstrated the applicability of formative assessment-based teaching in mathematics lessons. As far as the results are concerned, it is possible to say that appropriate formative assessment tools, which enabled the teacher to obtain information about the student faster, allowed the students to recognize their own success and improve their efficiency regarding success. The Teacher's Guide Booklet (MEB, 2020) sent to schools by the Ministry of National Education in 2020 focused on the importance of formative assessment-based teaching practice in mathematics education.

In fact, even though the importance of formative evaluation in both national and international education policies has been emphasized so much in recent years, there are only few studies in the national literature to validate its applicability (Ozan, 2017). Even though the positive aspects of formative assessment on students are commonly recognized by researchers, it is clear that scientific evidence applied to students has been limited (Bennett, 2011). Considering this lack of scientific evidence, the present study has clearly demonstrated that the formative assessment-based teaching application, based on the MEB (2020) guidebook, contributes to the students' mathematics achievement.

Furthermore, it is recommended that formative assessment-based mathematics teaching, which is widely used internationally, should be offered as part of the necessary teacher training in all schools in Turkey, its implementation should be tracked down, and it should be prioritized in the mobilization of teaching mathematics.

Acknowledgments

This article was presented as a paper at EDUCongress 2022. At the same time, this study was generated from one part of Hilal OZCAN's unpublished master's thesis under the supervision of Prof. Dr. Aytac KURTULUS.

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**: Eskişehir Osmangazi University, 25.01.2022, 2022-02.

Authorship Contribution Statement

Hilal Ozcan: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Aytac Kurtulus:** Methodology, Supervision, and Validation.

Orcid

Hilal Ozcan https://orcid.org/0000-0002-7460-1488
Aytac Kurtulus https://orcid.org/0000-0003-2397-3510

REFERENCES

Agwagah, U.N.V., & Ezieke, E.N. (2023). Effect of formative assessment practice on students' academic achievement in upper basic mathematics in Enugu East LGA, Enugu State, Nigeria. *Godfrey Okoye University International Journal of Education*, *3*(1), 1-10.

Aksu, Z. (2019). Pre-service mathematics teachers' pedagogical content knowledge regarding student mistakes on the subject of circle. *International Journal of Evaluation and Research in Education*, *8(*3), 440-445. https://files.eric.ed.gov/fulltext/EJ1232322.pdf

Amankonah, F.O. (2013). *K-8 teachers' self-efficacy beliefs for teaching mathematics* [Unpublished doctoral dissertation]. University of Nevada. https://search.proquest.com/docview/1496775105?accountid=15725

Arseven, Z. (2013). *İlköğretim matematik öğretmenlerinin 2005 ilköğretim matematik programında yer alan alternatif değerlendirme yaklaşımları uygulayabilme yeterliliklerinin incelenmesi* [*Examining the competencies of primary school mathematics teachers in applying alternative assessment approaches in the 2005 primary mathematics program*] [Master's thesis]. Uludağ University.

Austin Hurd, B.G. (2015). *How educators conduct formative assessment with middle school students in order to improve student achievement* [Unpublished doctoral dissertation]. Capella University.

Bennett, R.E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Burns, M.K., Klingbeil, D.A., & Ysseldyke, J. (2010). The effects of technology-enhanced formative evaluation on student performance on state accountability math tests. *Psychology in the Schools, 47*(6), 582-591. https://doi.org/10.1002/pits.20492

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö.E., Karadeniz, Ş., & Demirel, F. (2020). *Bilimsel araştırma yöntemleri* [Scientific research methods]. Pegem Academy.

Collins, N.M. (2012). *The impact of assessment for learning: Benefits and barriers to student achievement* [Unpublished doctoral dissertation]. Cardinal Stritch University.

Evirgen, O., & İkikardeş, N.Y. (2019). 7. sınıf matematik dersinde zorlanılan konulara ilişkin öğrenci görüşlerinin incelenmesi [Examining students' opinions on difficult topics in the 7th grade mathematics course]. *Journal of Balıkesir University Institute of Science and Technology, 21*(1), 416-435. https://doi.org/10.25092/baunfbed.548560

Fatima, I. (2022). *Examining the implementation of formative assessment practices in middle school mathematics: A basic qualitative study* [Doctoral dissertation]. Capella University.

Foster, D., & Poppers, A. (2009). Using formative assessment to drive learning. *The Silicon Valley Mathematics Initiative: The Noyce Foundation*. http://www.brjonesphd.com/uploads/1/6/9/4/16946150/using_formative_assessment_to_drive_learning_reduced.pdf

Görgüt, R.Ç. (2020). *Matematik öğretmen adaylarının matematiksel anlama boyutlarına yönelik etkinlik tasarım süreçlerinin incelenmesi: Çember ve daire* [*Examining pre-service mathematics teachers' activity design processes for mathematical understanding dimensions: Circle and circle*] [Unpublished doctoral thesis]. Ankara University.

Hiloma, K.A., & Briones, R.R. (2022). Learning activity sheets using interactive formative assessment and feedback mechanism for grade 9 mathematics. *International Journal of Research, 11*(7), 201-222. https://doi.org/10.5861/ijrse.2022.331

Kâhtalı, B.D., & Çelik, Ş. (2020). 2019 Türkçe öğretim programı'nda ölçme ve değerlendirme ile türkçe öğretmenlerinin ölçme ve değerlendirme araçlarını kullanma düzeyleri [Measurement and evaluation in the Turkish curriculum and Turkish teachers' levels of use of measurement and evaluation tools]. *Journal of Educational Theory and Practice Research, 6*(2), 237-244.

Kara, G. (2021). *Türkiye'de yayınlanan ortaokul matematik eğitimindeki kavram yanılgıları çalışmalarının incelenmesi* [*Examination of studies on misconceptions in secondary school mathematics education published in Turkey*] [Unpublished master's thesis]. Hacettepe University.

Karakuş, M., & Yeşilpınar, M. (2013). İlköğretim altıncı sınıf matematik dersinde uygulanan etkinliklerin ve ölçme-değerlendirme sürecinin incelenmesi: Bir durum çalışması [Examining the activities and measurement-evaluation process applied in the sixth-grade mathematics course of primary school: A case study]. *Pegem Journal of Education and Training, 3*(1), 35-54. https://dergipark.org.tr/en/pub/pegegog/issue/22585/241233

Kline, A.J. (2013). *Effects of formative assessment on middle school student achievement in mathematics and reading* [Unpublished doctoral dissertation]. North Carolina University.

Kültür, Y.Z. (2021). *Biçimlendirici değerlendirmenin ortaöğretim öğrencilerinin matematik ders başarısına ve tutumlarına etkisi* [*The effect of formative assessment on secondary school students' mathematics course achievement and attitudes*] [Doctoral thesis]. Çukurova University.

Martin, C., Mraz, M., & Polly, D. (2022). Examining elementary school teachers' perceptions of and use of formative assessment in mathematics. *International Electronic Journal of Elementary Education, 14*(3), 417-425. https://www.iejee.com/index.php/IEJEE/article/view/1764

Milli Eğitim Bakanlığı [MEB]. (2018). Matematik dersi öğretim programı (İlkokul ve Ortaokul 1,2, 3, 4, 5, 6, 7 ve 8. sınıflar) [Mathematics course curriculum (Primary and Secondary School 1,2, 3, 4, 5, 6, 7 and 8th grades)]. http://mufredat.meb.gov.tr/Dosyalar/201813017165445-MATE-MAT%C4%B0K%20%C3%96%C4%9ERET%C4%B0M%20PROGRAMI%202018v.pdf

Milli Eğitim Bakanlığı [MEB]. (2020). *Okul ve sınıf tabanlı değerlendirmeye dayalı öğretmen kapasitesinin güçlendirilmesi matematik dersi öğretmen rehber kitapçığı* [*Strengthening teacher capacity based on school and classroom-based assessment mathematics lesson teacher guidebook*]. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2020_08/26145631_Matematik.pdf

Milli Eğitim Bakanlığı [MEB]. (2022a). Ortaöğretim kurumlarına ilişkin merkezi sınav [Central examination for secondary education institutions]. Access: 03 September 2022, https://cdn.eba.gov.tr/icerik/2022/06/2022_LGS_rapor.pdf.

Milli Eğitim Bakanlığı [MEB]. (2022b). PISA 2018 Türkiye ön raporu [PISA 2018 Türkiye preliminary report]. http://www.meb.gov.tr/meb_iys_dosyalar/2019_12/03105347_PISA_2018_Turkiye_On_Raporu.pdf.

Miller, N. (2019). Formative assessment as a method to improve student performance in the sciences. *Honors Projects*, 461. https://scholarworks.bgsu.edu/honorsprojects/461

Ozan, C. (2017). *Biçimlendirici değerlendirmenin öğrencilerin akademik başarı, tutum ve öz düzenleme becerilerine etkisi* [*The effect of formative assessment on students' academic achievement, attitude and self-regulation skills*] [Doctoral thesis]. Atatürk University.

Önel, F., Dalkılınç, F., Özel, N., Deniz, Ş., Balkaya, T., & Birel, G. (2020). Ortaokul matematik öğretmenleri ölçme-değerlendirmeyi nasıl yapıyor? Bir durum çalışması [How do secondary school mathematics teachers conduct assessment and evaluation? A case study]. *Journal of Kastamonu Education, 28(*3), 1448-1459. https://doi.org/10.24106/kefdergi.4113

Tekin, E.G. (2010). *Matematik eğitiminde biçimlendirici değerlendirmenin etkisi* [*The effect of formative assessment in mathematics education*] [Published master's thesis]. Marmara University.

Van de Walle, J.A., Karp, K.S., Bay-Williams, J.W., & Durmuş, S. (Ed). (2019). *İlkokul ve ortaokul matematiği gelişimsel yaklaşımla öğretim* [*Teaching primary and secondary school mathematics with a developmental approach*]. Nobel Publications.

Wafubwa, R.N., & Csíkos, C. (2022). Impact of formative assessment instructional approach on students' mathematics achievement and their metacognitive awareness. *International Journal of Instruction, 15*(2), 119-138. https://eric.ed.gov/?id=EJ1341771

Yıldırım, A., & Şimşek A. (2021). *Sosyal bilimlerde nitel araştırma yöntemleri* [Qualitative research methods in the social sciences]. Seckin Publishing.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 765–780

https://doi.org/10.21449/ijate.1250358

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Purification procedures used for the detection of gender DIF: Item bias in a foreign language test

Serap Buyukkidik[1,*]

[1]Sinop University, Faculty of Education, Department of Educational Sciences, Sinop, Türkiye

ARTICLE HISTORY

Received: Feb. 12, 2023
Revised: Oct. 21, 2023
Accepted: Nov. 03, 2023

Keywords:

Mantel-Haenszel,
Simultaneous item bias test,
Lord's chi-square,
Raju's unsigned area,
Bias,
Item purification.

Abstract: In the current study, differential item functioning (DIF) detection using real data was conducted with the application of "Mantel-Haenszel (MH)", "Simultaneous item bias test (SIBTEST)", "Lord's chi-square", and "Raju's area" methods, both when item purification was carried out and when item purification was not. After detecting gender-related DIF, expert opinions were obtained for a bias study since it is important to conduct gender bias research in the English test. Additionally, in the relevant literature, there were some DIF studies, but not completely similar bias studies. The sample of the research consisted of 7,389 students who took the "Transition from Primary to Secondary Education Exam (TPSEE, referred to as "TEOG" in Turkish)" administered in April 2017. When gender-related DIF analysis was performed with the aforementioned four methods, the results were found to differ partially. DIF analysis results differed in different conditions based on whether item purification was performed or not. Furthermore, the detection of DIF was indicative of potential bias. In the second stage of the study, the opinions of seven experts were sought for item 11, for which DIF was detected at least at B level based on MH, SIBTEST. As a result of expert opinion, it was established that there was no bias based on gender in any of the items in the English test. It is advised that akin bias studies be carried out to enable test developers to be aware of characteristics that may result in item bias and construct unbiased items.

1. INTRODUCTION

There is much research on gender differences in foreign language testing. Gender differences in the acquisition of a second language are controversial and emerge as a prominent topic in the literature (Llach & Gallego, 2012). However, the main question to be asked in such research is "Are these differences due to the real differences in the measured trait of different gendered individuals?" or "Do these differences stem from item bias?". These questions have rarely been asked by researchers who conduct gender differences in achievement research.

The fairness and validity of the test are threatened in a test consisting of biased items (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014). People differ in terms of many demographic variables such as culture, gender, language, and ethnicity. The educational and

*CONTACT: Serap BUYUKKIDIK ✉ sbuyukkidik@gmail.com Sinop University, Faculty of Education, Department of Educational Sciences, Sinop, Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

psychological assessment of this wide variety of people must be carried out with the same precision and fairness across groups, regardless of their irrelevant characteristics (gender, etc.) (Sireci & Rios, 2013). Item bias or differential item functioning (DIF) affects test fairness (Khalid & Glas, 2014). DIF and bias are two separate concepts (AERA, APA, & NCME, 2014). DIF shows the difference in the probability of individuals at the same ability levels responding correctly to the item and differentiation as a function of group membership (Hambleton & Rogers, 1989; Holland & Wainer, 1993; Camilli & Shepard, 1994; Zumbo, 1999; AERA, APA, & NCME, 2014). There are two reasons for detecting DIF: item bias and item effect, which are the real differences between subgroups (Camilli & Shepard, 1994). In other words, the detection of DIF is not always an indicator of bias (AERA, APA, & NCME, 2014). Bias is the systematic error in the item and test performances of individuals in different subgroups depending on the subgroup they belong to (Osterlind, 1983; Crocker & Algina, 1986; Camilli & Shepard, 1994; Zumbo, 1999; AERA, APA, & NCME, 2014). In bias studies, DIF analyses are performed at the first stage, and then, important reasons for the item bias are found by the expert opinion method. While the detection of DIF is a statistical process, the detection of item bias is a conceptual process based on interpretation (Camilli & Shepard, 1994; Zumbo, 1999; Wiberg, 2007). Item bias studies date back to Alfred Binet's test of mental capacity in 1910 (Camilli & Shephard, 1994).

Although many studies detect DIF today, the number of bias studies is quite limited despite its long history. In bias studies, it is seen that DIF studies are conducted without item purification generally (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya, 2012; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019). However, no bias study has encountered real data that deal with how the results differ with and without item purification in the detection of DIF. In this respect, there arises a need to conduct such bias studies in the literature.

1.1. Differential Item Functioning and Item Purification

Although the first studies on DIF were conducted by Cardall and Coffman (1964) and Angoff and Ford (1973) in the 20th century (Holland & Thayer, 1986), fairness in educational and psychological measures in the 21st century is a current issue that researchers still give importance (Sireci & Rios, 2013). There are many methods for detecting DIF and estimating its size. Some of these methods include Mantel-Haenszel, SIBTEST, IRT methods, standardization, chi-square, Likelihood ratio test, Logistic Regression, b parameter indices, probability differences indices, IRT Likelihood Ratio Test (LRT), general IRT-LR, log-linear models, and Lord's chi-squared test (Wiberg, 2007). Since these DIF detection methods are based on different statistical bases, detecting DIF using different methods may lead to different results (Çepni & Kelecioğlu, 2021; Bakan Kalaycıoğlu, 2022). Regardless of which DIF detection method is used, there are two groups in DIF analyses, the “focal group and the reference group”, and different functionalization between these groups is considered.

DIF detection methods can be classified in terms of “parametric vs. non-parametric”, “matching variable: observed vs. latent”, “dichotomously vs. polytomously”, “measure and/or test of DIF”, “uniform vs. non-uniform DIF”, “handle the cut-off score or not”, “sample size” (Wiberg, 2007). In addition, DIF is divided into uniform and non-uniform. If the probability of answering an item correctly is in favor of a group at all skill levels, it is said to be uniform DIF. If the probability of answering an item correctly is in favor of a different group at different skill levels, it is said to be non-uniform DIF (Camilli & Shepard, 1994; Zumbo, 1999). For example, while uniform DIF is detected in the MH method, non-uniform DIF can be detected in the Crossing-SIBTEST (CSIBTEST) developed in addition to the SIBTEST (Wiberg, 2007).

Using different DIF detection methods can affect results. Another factor affecting the differentiation of results in the detection of DIF is item purification The indication that the element for which the DIF is not detected means that the DIF detected in that element causes a type 1 error. Item purification is an iterative process used to control the error rate and increase the power and precision of the results (Khalid & Glas, 2014). According to Lord (1980), eliminating DIF items in iterative and multiple stages purifies test scores and reduces power and Type 1 error. Fidalgo, et al. (2000) discovered that different purification types (three amounts of DIF "(10%, 15%, and 30% of DIF-items), three test lengths (20, 40, and 60 items)" under different simulation conditions (single-stage, two-stage, and iterative) investigated the effect of the MH DIF detection method on performance. Based upon the findings of their research, they stated that the two-stage purification process was more effective than the one-stage purification process. Wang and Su (2004) suggested that two-stage and iterative purification could be safely used to reduce the inflated Type 1 error as a result of the Monte Carlo simulation study. When the related studies were examined, some studies suggested the iterative purification process (Lord, 1980; Fidalgo et al., 2000; Wang & Su, 2004; Khalid & Glas, 2014), but some of the research showed that purification does not improve the detection of DIF (Magis & Facon, 2013), indicating that there is no definitive conclusion. In this respect, it is important to conduct studies that reveal how DIF detection is affected when purification is performed and when it is not. When the purification studies were examined, it was seen that there were mostly simulation studies for the MH method (Wang & Su, 2004; Fidalgo et al., 2000). Studies comparing DIF results with and without purification on real data were quite limited (e.g., Özdemir, 2015; Tunc et al., 2018; Soysal & Yılmaz Koğar, 2021).

There have been many studies on DIF detection in the literature. However, most of these studies compared at least two methods (e.g., Emily et al., 2021; Soysal & Yılmaz Koğar, 2021). Emily, et al. (2021) performed Monte Carlo simulation and examined Lord's Chi-square (LC), LRT, and MH DIF detection methods in terms of type 1 error and found that the MH method had the best performance in terms of type 1 error. Soysal and Yılmaz Koğar (2021), on the other hand, determined DIF based on Lord's χ^2 and Raju's unsigned area methods. It was found that a limited number of bias studies are carried out in large-scale or national examinations (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya, 2012; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019). In the bias studies conducted the effect of item purification was considered. In this study, we aimed to conduct comparative DIF analyses based on MH, SIBTEST, Lord's χ^2, and Raju's unsigned area methods in real data set when item purification was or was not performed, and to conduct bias studies by obtaining expert opinions on items containing at least moderate DIF in at least two methods. In this respect, the study differed from similar studies. It was thought that the research would contribute to the literature. The reason for using Lord's χ^2 and Raju's unsigned area methods in this research is that although there have been DIF studies using these two DIF detection methods with item purification and without item purification (e.g., Özdemir, 2015; Tunc, et al., 2018; Soysal & Yılmaz Koğar, 2021), there are no bias studies in the literature that consider purification. In addition, Tunc, et al. (2018) found that these two methods were the most sensitive in the purification process, which was effective in the selection of these methods. In addition, MH and SIBTEST are among the most used DIF detection methods in the literature. That's why these two methods were included in the study.

This specific study aimed to investigate whether the "Transition from Primary to Secondary Education Exam (TPSEE, referred to as "TEOG" in Turkish)", which was administered in April 2017 showed a gender bias. For this purpose, DIF detection was first carried out and if no purification was carried out, a DIF detection was carried out. Then, expert opinion was sought for the bias study. The study attempted to answer the following questions within the scope of the research:

1) In the TPSEE 2017 English test without purification, which items were gender-related-DIF detected based on the MH, SIBTEST and Crossing-SIBTEST, Lord's χ^2, and Raju's unsigned area methods?

2) When purification is performed, in which items of TPSEE 2017 English test, gender-related DIF detected based on MH, SIBTEST and Crossing-SIBTEST, Lord's χ^2, and Raju's unsigned area methods?

3) Which items in the TPSEE 2017 English test are gender biased according to expert opinion?

2. METHOD

2.1. Participants

The sample of the research consisted of 7,389 8th-grade students who took the TPSEE administered in April 2017. While 3,606 of these students were females, the other 3,783 students were males. Table 1 shows descriptive statistics for the total group, the focal group, and the reference group. In this study, females were treated as the reference group and males as the focal group. While the mean of the reference group is 13.58, the mean of the focal group is 11.28. The average test score of the female students is higher than that of the male students.

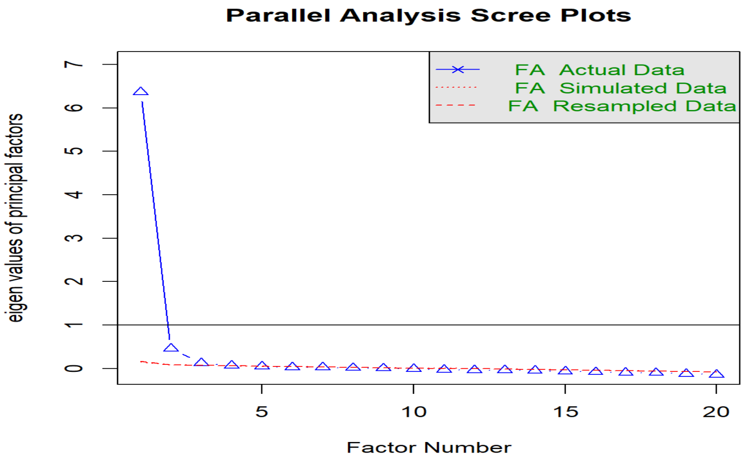
Table 1. *Descriptive statistics for reference and focal group.*

	n	Min	Max	Mean	Median	SD
Total	7389	0.00	20.00	12.40	12.00	5.55
Reference group-Female (0)	3606	0.00	20.00	13.58	14.00	5.35
Focal group- Male (1)	3783	0.00	20.00	11.28	10.00	5.51

2.2. Instrument

TPSEE is a national high school entrance exam for 8th-grade students, which was administered by the Turkish Ministry of National Education from 2013 to 2017. The exam consisted of six subtests including maths, science and technology, Turkish, English (as a foreign language), Turkish Revolutionary History and Kemalism, and religion, culture, and ethics. The data collection instrument was an English (foreign language) subtest (Booklet A) of TPSEE consisting of 20 multiple-choice items. The psychometric properties of the English subtest are discussed in this section. For the DIF analysis based on IRT, l, the unidimensionality assumption was first tested. As a result of the parallel analysis based on tetrachoric correlation, the English test was found to have a uni-dimensional structure (see Figure 1).

Figure 1. *Parallel analysis scree plots.*



When model data fit indices were examined, it was found that the data were suitable for one-dimensional structure ($\chi^2_{(170)}$ = 3760.91; p = 0, RMSEA = 0.064, 90% CI [0.062, 0.065], TLI = 0.936). Factor 1 explains 49.95% of the variance. The factor loadings ranged from 0.47 (item 16) to 0.82 (item 18). The reliability of the measurements was good enough (KR-20 = 0.90, marginal reliability for 3 PL= 0.82).

When performing the item analysis in IRT, the model data fits in 1 PL, 2 PL, and 3 PL were examined. The model data fits for each model were treated according to five criteria. The model data fits for each model are presented in Table 2.

Table 2. *Model-Data Fit.*

Model	AIC	AICc	BIC	SABIC	logLik
1PL	151026.19	151026.315	151171.253	151104.519	-75492.095
2PL	149191.191	149191.637	149467.501	149340.389	-74555.595
3PL	146169.156	146170.154	146583.62	146392.953	-73024.578
BEST	3PL	3PL	3PL	3PL	3PL

The 3 PL had the best model-data fit based on all five criteria (Akaike information criterion (AIC), Corrected AIC, Bayesian information criterion (BIC), sample-size adjusted BIC (SABIC), and likelihood ratio test (logLik)). Item parameters were obtained on the basis of 3 PL. Considering the discrimination parameter "*a*", it was seen that item 3 had the lowest discrimination (a_3=1.77), whereas item 17 had the highest discrimination (a_{17}=6.7). When the "*b*" item difficulty parameters were examined on the basis of 3 PL, item 13 had the lowest difficulty parameter (b_{13}=-1.25), while item 5 had the highest *b* parameter (b_5=0.68). When the "*c*" values or guessing parameters were examined, the item with the highest probability of answering by luck was item 16 (c_{16}=0.41), while item 3 had the lowest c parameter (c_3=0.01).

2.3. Analysis of Data

First, the missing data (111 student responses) were removed from the data set. The analysis of the data then began. In the analysis of the data, the construct validity and reliability proofs were collected in the first stage. In the second stage, unidimensionality and local independence from IRT assumptions were tested. Parallel analysis based on tetrachoric correlation was performed for unidimensionality (see Figure 1). For local independence, Yen's Q_3 index was examined and binary values were found below 0.20 in this research. To calculate the item parameters, analyses were performed according to 1 PL, 2 PL, and 3 PL. As the best model fit was achieved in the 3 PL, item parameters were considered according to the 3 PL. "irtGUI" package in R programming language (2021) was used for IRT assumptions (Yen's Q_3 index and parallel analysis) and for estimating marginal reliability.

In this study, DIF analyses based on 3PL were performed in IRT-based methods. After providing the assumptions, DIF analyses were performed with MH, SIBTEST, Lord's chi-square, and IRT Raju's unsigned area test. DIF analyses and parameter estimation based on IRT were performed using ShinyItemAnalysis (Martinkova & Drabinova, 2018) based on "mirt" (for item parameter estimation), "difR", and "ltm" (for DIF analysis) packages in R programming language (2021).

The code was generated for each method to indicate whether each method (MH, SIBTEST and Crossing-SIBTEST, Lord's χ^2, and Raju's unsigned area) underwent purification separately, and the items found when purification was performed (1) and not performed (0) for 20 items. The agreement coefficient was calculated using ReCal (Freelon, 2013) to assess the agreement between the results.

In the final stage, a bias study was conducted by taking expert opinions on an item containing moderate DIF. Some characteristics of the experts whose opinions were used for the bias study are shown in Table 3.

Table 3. *Information about experts.*

Id	Gender	Experience	Field	Id	Gender	Experience	Field
E1	Male	9 years	Measurement and Evaluation (MS, Ph.D.), Language Teaching (Undergraduate)	E5	Female	9 years	Measurement and Evaluation (MS, Ph.D.), Language Teaching (Undergraduate)
E2	Female	21 years	English Literature (Ph.D.) English Language Teaching (Undergraduate, MA)	E6	Female	11 years	English Language Teaching (Undergraduate, MS, Ph.D.)
E3	Female	15 years	English Literature (MA, Ph.D.) English Language Teaching (Undergraduate)	E7	Male	13 years	English Language Teaching (Undergraduate, MS, Ph.D.)
E4	Male	15 years	English Language Teaching (Undergraduate, MS, Ph.D.)				

Table 3 shows that a total of seven experts (two measurement and evaluation specialists and five English language teaching specialists) participated in the research. Four of these experts were lecturers in higher education who had completed their Ph.D. in English Language Teaching. While one of the other three experts was completing a Ph.D. in English Literature, two of them graduated from English language teaching and had completed their master's degree in measurement and evaluation and they were currently continuing their Ph.D. education. All experts had a master's degree in their field and had at least 9 years of experience. Descriptive data analysis was used for the qualitative part of the research in the bias study. The DIF detection methods used in the study are as follows.

2.3.1. *The Mantel-Haenszel test*

The Mantel and Haenszel test was developed by healthcare workers Mantel and Haenszel (1959), and its use in the detection of DIF is based on the work of Holland and Thayer (1986). The MH method uses the 2x2xK contingency table (Holland & Thayer, 1986). MH is a non-parametric, uniform DIF detection method based on classical test theory (CTT) that can be tested on polytomous and binary data (Wiberg, 2007). Zieky (1993) established reference ranges for the determination of DIF levels taking IΔMHI into account. When $|\Delta MH|<1$ there is no DIF or A-level ie negligible level. When it is in the range of $1\leq|\Delta MH|<1.5$, moderate (level B) DIF is detected. In the range of $|\Delta MH|\geq1.5$, a high-level (C level) DIF is detected.

2.3.2. *Simultaneous item bias test*

The SIBTEST for the detection of uniform DIF was developed by Shealy and Stout (1993) based on the standardization method. The basis of the CSIBTEST method developed to detect non-uniform DIF is based on the work of Li and Stout (1996), followed by Chalmers (2018). β values are obtained in the SIBTEST. When β is negative, the focal group is advantaged, and when β is positive, the reference group is advantaged (Gierl & Khaliq, 2001). In addition, Roussos and Stout (1996) suggested DIF classification according to the magnitude of the β value. When β is | β |<0.059, there is no DIF or A-level ie negligible level. When β is in the range of 0.059≤| β |<0.088, moderate (B-level) DIF is detected. If β is in the range of | β |≥0.088, a high level (C level) DIF is detected. SIBTEST is a method that allows the analysis of non-parametric binary and polytomous data (Wiberg, 2007).

2.3.3. *Lord's chi-square test*

Lord's chi-squared test is an IRT-based parametric DIF detection method based on Lord's work in 1980. It is an extended version of Lord's chi-squared test, the test of the b difference method, including the distinctiveness parameter (Lord, 1980). The Lord's chi-squared test allows the detection of DIF by taking into account the item parameters and the difference between the groups. Lord's chi-squared test is a method for detecting both uniform and non-uniform DIF that can be used in binary data (Wiberg, 2007). However, this method does not measure the size of DIF; it only tests for the presence of a DIF item (Wiberg, 2007).

2.3.4. *Raju's area method*

Raju's area method is based on Raju's work in 1988 and 1990. This method is a parametric method based on IRT. The logic of this method is based on the area between the item characteristic curves of the focal and reference groups in the signed area (Raju, 1988). In the null hypothesis, this area is equal to zero. Z statistics are used for this purpose. In recent studies, the unsigned area method is used. Raju's unsigned area is used to detect non-uniform DIF. Raju's unsigned area method is computed from the difference between the difficulty and discrimination parameters (Raju, 1988). One of the major problems with Raju's area methods (both signed and unsigned area) is their limitation for 3PL estimation. Raju (1988, 1990) showed that the area between two item response functions is infinite when the lower asymptotes are not equal. Raju (1988, 1990) suggested that equal or fixed c-parameters should be used for this situation. The guessing parameter *c* is estimated from the entire dataset and is considered fixed in the present study under the 3PL model.

3. RESULTS

In this section, the results of MH, SIBTEST, Lord's chi-square, and Raju's unsigned area method are given in terms of whether or not item purification was performed. DIF results are given for the 20-item English test.

3.1. DIF Results When No Item Purification was Performed

Table 4 shows the results of DIF analysis without item purification based on the MH method.

Table 4. *DIF results based on MH.*

Item	*MH*(χ^2)	*p*-value	αMH	ΔMH	DIF Level	Advantage group
item1	7.26	0.01	0.84	0.40	A	Male
item2	4.55	0.03	1.14	-0.30	A	Female
item3	6.89	0.01	1.20	-0.42	A	Female
item4	7.14	0.01	0.84	0.42	A	Male
item5	17.44	0.00	0.77	0.62	A	Male
item6	8.51	0.00	1.21	-0.45	A	Female
item7	11.80	0.00	1.28	-0.59	A	Female
item8	0.28	0.60	0.96	0.08	-	-
item9	2.79	0.10	1.11	-0.24	-	-
item10	0.28	0.60	1.04	-0.09	-	-
item11	**55.69**	**0.00**	**1.73**	**-1.28**	**B**	**Female**

item12	2.64	0.10	0.90	0.26	-	-
item13	4.85	0.03	0.82	0.47	A	Male
item14	0.14	0.71	1.02	-0.05	-	-
item15	0.03	0.87	1.01	-0.03	-	-
item16	6.21	0.01	1.15	-0.32	A	Female
item17	9.02	0.00	0.84	0.41	A	Male
item18	15.92	0.00	0.76	0.64	A	Male
item 19	0.09	0.76	1.02	-0.05	-	-
item 20	8.26	0.00	0.83	0.42	A	Male

When item purification was not performed in the MH method, DIF was detected in item 1, item 2, item 3, item 4, item 5, item 6, item 7, item 11, item 13, item 16, item 17, item 18, and item 20. Item 1, item 4, item 5, item 13, item 17, item 18, and item 20 contained negligible DIF in favor of male students. Item 2, item 3, item 6, item 7, item 11, and item 16 contained DIF in favor of females. Only a moderate DIF level was detected in item 11, e.i., a B level of DIF. A-level DIF was detected in all other items containing DIF.

Table 5 shows the results of DIF analysis without item purification based on SIBTEST and CSIBTEST methods.

Table 5. *DIF results based on SIBTEST and Crossing-SIBTEST.*

Item	β_{uni}	β_{cro}	X_{uni}^2	χ_{cro}^2	p-value$_{uni}$	p-value$_{cro}$	Uniform/Non-uniform	DIF Level	Advantage group
item1	-0.03		5.34		0.02		Uniform	A	Male
item2		0.03		8.36		0.02	Non-uniform	A	-
item3	0.04		11.68		0.00		Uniform	A	Female
item4	-0.03		6.28		0.01		Uniform	A	Male
item5	-0.05		16.98		0.00		Uniform	A	Male
item6	0.03		8.01		0.00		Uniform	A	Female
item7	0.04		15.46		0.00		Uniform	A	Female
item8	-0.00	0.00	0.08	0.08	0.77	0.77	NO DIF	-	-
item9	0.03		5.22		0.02		Uniform	A	Female
item10	0.01	0.02	0.93	2.52	0.33	0.28	NO DIF	-	-
item11	**0.08**		**60.55**		**0.00**		**Uniform**	**B**	**Female**
item12	-0.01	0.01	1.70	1.70	0.19	0.19	NO DIF	-	-
item13	-0.02	0.02	3.82	3.82	0.05	0.05	NO DIF	-	-
item14	0.01	0.03	0.62	5.47	0.43	0.06	NO DIF	-	-
item15	0.01	0.01	0.27	1.02	0.61	0.60	NO DIF	-	-
item16	0.03		6.95		0.01		Uniform	A	Female
item17		0.03		9.85		0.01	Non-uniform	A	-
item18	-0.04		18.75		0.00		Uniform	A	Male
item19	0.00	0.01	0.02	1.06	0.89	0.59	NO DIF	-	-
item20		0.03		11.72		0.00	Non-uniform	A	-

Table 5 shows that uniform DIF was detected in item 1, item 3, item 4, item 5, item 6, item 7, item 9, item 11, item 16, and item 18. Non-uniform DIF was detected in item 2, item 17, and item 20. Only item 11 contains DIF at the B level in the English test, while DIF at the A level was detected in the other items containing DIF. Item 1, item 4, item 5, and item 18 contained

DIF in favor of males. DIF was detected in favor of females in item 3, item 6, item 7, item 9, item 11, and item 16.

Table 6 shows the findings of DIF analysis without item purification according to Lord's χ^2 and Raju's unsigned area methods.

Table 6. *DIF results based on Lord's χ^2 ve Raju's unsigned area.*

Item	Lord's χ^2	*p*-value	Raju's Z	*p*-value	Item	Lord's χ^2	*p*-value	Raju's Z	*p*-value
item1	3.52	0.17	-1.87	0.06	**item11**	**28.37**	**0.00**	**-4.77**	**0.00**
item2	22.88	0.00	-4.44	0.00	item12	2.36	0.31	1.56	0.12
item3	0.74	0.69	0.81	0.42	item13	20.50	0.00	2.86	0.00
item4	7.97	0.02	-2.86	0.00	item14	0.04	0.98	-0.19	0.85
item5	2.82	0.24	-1.68	0.09	item15	0.58	0.75	0.67	0.51
item6	9.27	0.01	-2.81	0.00	item16	11.31	0.00	3.41	0.00
item7	3.24	0.20	1.78	0.07	item17	13.23	0.00	-2.96	0.00
item8	0.01	1.00	-0.09	0.93	item18	12.86	0.00	-3.71	0.00
item9	6.42	0.04	2.53	0.01	item19	4.65	0.10	-2.16	0.03
item10	3.41	0.18	1.78	0.08	item20	14.70	0.00	3.88	0.00

Table 6 shows that DIF based on both Lord's χ^2 and Raju's unsigned area methods was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, item 17, item 18, and item 20. DIF was detected in item 19 based on only Raju's unsigned area method.

It was seen that DIF was detected in item 2, item 4, item 6, item 11, item 16, item 17, item 18, and item 20 based on four methods when item purification was not performed. Only item 11 contained a moderate DIF in favor of girls. Other DIF-containing items contained negligible levels of DIF.

3.2. DIF Results When Item Purification was Performed

Table 7 shows the results of the DIF analysis based on the MH method when item purification was performed. When item purification was performed based on the MH method, it was detected that DIF was detected in item 1, item 3, item 4, item 5, item 7, item 11, item 12, item 13, item 17, item 18, and item 20. Item 1, item 4, item 5, item 12, item 13, item 17, item 18, and item 20 contained DIF in favor of males. Item 3, item 7, and item 11 contained DIF in favor of females. Only item 11 contained moderate DIF, while the rest of the DIF-detected items contained negligible (A-level) DIF.

Table 7. *DIF analysis results based on MH when item purification was performed.*

Item	*MH*(χ^2)	*p*-value	αMH	ΔMH	DIF Level	Advantage group
item1	12.33	0.00	0.79	0.54	A	Male
item2	1.12	0.29	1.07	-0.16	-	-
item3	4.26	0.04	1.16	-0.34	A	Female
item4	13.05	0.00	0.78	0.59	A	Male
item5	35.54	0.00	0.68	0.90	A	Male
item6	2.05	0.15	1.10	-0.23	-	-
item7	7.41	0.01	1.23	-0.48	A	Female
item8	3.39	0.07	0.88	0.29	-	-
item9	0.00	1.00	1.00	-0.00	-	-
item10	0.00	1.00	1.00	-0.01	-	-

item11	**40.09**	**0.00**	**1.61**	**-1.12**	**B**	**Female**
item12	4.73	0.03	0.86	0.36	A	Male
item13	4.33	0.04	0.83	0.45	A	Male
item14	0.75	0.39	0.95	0.13	-	-
item15	0.01	0.93	0.99	0.02	-	-
item16	1.42	0.23	1.07	-0.16	-	-
item17	21.76	0.00	0.76	0.65	A	Male
item18	27.42	0.00	0.69	0.86	A	Male
item19	1.28	0.26	0.93	0.18	-	-
item20	12.55	0.00	0.80	0.54	A	Male

In Table 8, DIF analysis findings in SIBTEST and CSIBTEST are given when item purification was performed. When item purification was performed, Table 8 shows that uniform DIF was detected in item 2, item 3, item 5, item 6, item 7, item 9, item 10, item 11, item 15, item 16, and item 18. Only item 11 contained DIF at the C level, while other DIF-detected items contained DIF at the A level. Item 5 and item 18 contained DIF in favor of males. DIF was detected in favor of females in item 2, item 3, item 6, item 7, item 9, item 10, item 11, item 15, and item 16.

Table 8. *DIF analysis results based on SIBTEST and Crossing-SIBTEST when item purification was performed.*

Item	β_{uni}	β_{cro}	X_{uni}^2	χ_{cro}^2	p-value$_{uni}$	p-value$_{cro}$	Uniform/ Non-uniform	DIF Level	Advantage group
item1	-0.01	0.02	0.37	2.23	0.54	0.33	NO DIF	-	-
item2	0.03		7.39		0.01		Uniform	A	Female
item3	0.06		25.38		0.00		Uniform	A	Female
item4	-0.00	0.02	0.02	3.24	0.89	0.20	NO DIF	-	-
item5	-0.03		8.62		0.00		Uniform	A	Male
item6	0.05		18.61		0.00		Uniform	A	Female
item7	0.06		30.44		0.00		Uniform	A	Female
item8	0.02	0.02	2.84	2.84	0.09	0.09	NO DIF	-	-
item9	0.05		17.31		0.00		Uniform	A	Female
item10	0.04		10.34		0.00		Uniform	A	Female
item11	**0.10**		**82.43**		**0.00**		**Uniform**	**C**	**Female**
item12	0.02	0.02	2.56	2.56	0.11	0.11	NO DIF	-	-
item13	0.00	0.00	0.08	0.08	0.77	0.77	NO DIF	-	-
item14	0.02	0.02	3.60	3.60	0.06	0.06	NO DIF	-	-
item15	0.03		7.72		0.01		Uniform	A	Female
item16	0.05		11.64		0.00		Uniform	A	Female
item17	-0.02	0.03	3.47	5.45	0.06	0.07	NO DIF	-	-
item18	-0.02		3.95		0.05		Uniform	A	Male
item19	0.02	0.02	2.82	2.82	0.09	0.09	NO DIF	-	-
item20	-0.00	0.02	0.01	3.13	0.94	0.21	NO DIF	-	-

Table 9 shows the results of DIF analysis when item purification was performed in Lord's χ^2 and Raju's unsigned area methods.

Table 9. *DIF analysis results based on Lord's χ^2 ve Raju's unsigned area when item purification was performed.*

Item	Lord's χ^2	*p*-value	Raju's Z	*p*-value	item	Lord's χ^2	*p*-value	Raju's Z	*p*-value
item1	2.80	0.25	-1.74	0.08	**item11**	**31.49**	**0.00**	**-4.36**	**0.00**
item2	42.31	0.00	-5.95	0.00	item12	0.14	0.93	-0.56	0.57
item3	1.49	0.47	-1.06	0.29	item13	28.60	0.00	3.84	0.00
item4	6.03	0.05	-2.73	0.01	item14	4.14	0.13	-1.85	0.06
item5	3.16	0.21	-1.65	0.10	item15	0.92	0.63	-0.94	0.35
item6	25.08	0.00	-4.42	0.00	item16	23.69	0.00	4.66	0.00
item7	1.06	0.59	0.84	0.40	item17	1.33	0.51	-1.52	0.13
item8	3.85	0.15	-1.82	0.07	item18	16.53	0.00	-4.40	0.00
item9	14.56	0.00	3.64	0.00	item19	18.61	0.00	-4.29	0.00
item10	0.27	0.87	0.43	0.67	item20	5.10	0.08	2.20	0.03

Table 9 shows that when item purification was performed, DIF based on both Lord's χ^2 and Raju's unsigned area methods was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, item 18, and item 19. Item 20 only contained DIF based on Raju's unsigned area method.

3.3. Comparison of DIF Results

Table 10 contained items for which DIF was detected in different conditions.

Table 10. *Comparison of DIF methods.*

Mantel-Haenszel test		SIBTEST		IRT Lord's χ^2		IRT Raju's area test	
Without item purification	With item purification	Without item purification	With item purification	Without item purification	With item purification	Without item purification	With item purification
item1, item2, item3, item4, item5, item6, item7, item11, item13, item16, item17, item18, item20	item1, item3, item4, item5, item7, item11, item12, item13, item17, item18, item20	item1, item2, item3, item4, item5, item6, item7, item9, item11, item16, item17, item18, item20	item2, item3, item5, item6, item7, item9, item10, item11, item15, item16, item18	item2, item4, item6, item9, item11, item13, item16, item17, item18, item20	item2, item4, item6, item9, item11, item13, item16, item18, item19	item2, item4, item6, item9, item11, item13, item16, item17, item18, item19, item20	item2, item4, item6, item9, item11, item13, item16, item18, item19, item20
13 items	11 items	13 items	11 items	10 items	9 items	11 items	10 items

It was found that the results of DIF analysis differed partially with or without item purification. For example, item 20 contained DIF based on four methods when item purification was not performed, while it contained DIF based on the MH and Raju's unsigned area method when item purification was performed. When item purification was not performed, DIF was detected in item 17 based on four methods, whereas when item purification was performed, it contained

DIF only based on the MH method. Whether item purification was performed or not, DIF was detected in item 3, item 5, item 7, item 11, and item 18 based on SIBTEST and MH methods. Whether item purification was performed or not, DIF was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, and item 18 based on the Lord's chi-square method. Based on Raju's unsigned area test, DIF was detected in item 2, item 4, item 6, item 9, item 11, item 13, item 16, item 18, item 19, and item 20 in both cases. When item purification was not performed, DIF was detected in 13 items based on SIBTEST and MH methods. Based on the Lord's chi-square method, DIF was detected in 10 items and 11 items based on Raju's unsigned area method. When item purification was performed, DIF was detected in 11 items based on the SIBTEST and MH methods. Based on Lord's chi-square method, DIF was detected in 9 items. Based on Raju's unsigned area method, DIF was detected in 10 items. It was found that the results differed partially from method to method and according to the condition of item purification.

For the MH method, the average pairwise percent agreement was found to be 80% whether or not purification was performed. For the SIBTEST method, the average pairwise percent agreement was found to be 70% in the same condition. For the Lord's chi-square method, the agreement was 85%, and for Raju's unsigned area method, it was 95% in the same condition. The highest level of agreement was found for Raju's unsigned area method in the condition where purification was both performed and not performed.

3.4. Expert Opinions for Bias Study

Table 11 summarizes the opinions of the experts on item 11, for which at least a moderate DIF level was detected as a result of the DIF analysis. Looking at Table 11, it can be seen that there were 6 experts (85.71%) who stated that item 11 was not biased, while one expert stated that it was biased. In the face-to-face interviews with Expert 2, she expressed that she was not sure about the bias of the item.

Table 11. *Expert opinions for the item 11.*

Expert number	Decisions	Expert number	Decisions
E1	No bias	E5	No bias
E2	Bias	E6	No bias
E3	No bias	E7	No bias
E4	No bias	Total	6 no bias (%85,71), 1 bias (%14,19)

Expert 2' explanation was as follows: "Expression types and patterns in the content of the visual and item used together to cause a bias towards gender… It is known that reading texts and writing activities vary according to gender in terms of preferences. The interests of male and female students may differ depending on the genre. For example, in the case of reading activities, it has been found in the literature that "males prefer to read texts for a purpose such as gaining knowledge and learning how to do something" or, in written activities, "females are much more interested in the activity of writing a letter to a pen pal than male students" ... As a genre in which feelings and thoughts are conveyed, the letter is a communication and bonding tool for female students. From the first years of education, it can be observed that girls write to each other in short notes or in letter format. For these reasons, the expression tested in the question creates more familiarity for girls as a type of writing and can make it easier for them to notice the details in the image. On the other hand, the fact that boys have more access to technological communication tools in terms of opportunities causes them to spend more time with such tools. Therefore, the expressions in e-mails, voicemails, and text messages in the options may attract their attention more as a distraction and may cause them to turn to a wrong answer without fully evaluating the image."

Expert 4 (E4), on the other hand, expressed a different opinion and said, "When the root of the item in the test and the options were examined, no situation that could cause a bias in terms of gender was observed. ". When we asked, “What is your opinion about the item containing DIF in favor of girls/DIF source?” questions, he said “The question is a question that can be answered correctly according to the detail in the image. The fact that female students are more successful in recognizing details than male students may be a source of DIF. And this points to the real difference in students' ability levels, not bias.”. When asked about the reason for this situation, he said, “The question is a question where the correct answer can be found according to the detail in the image. The fact that female students are more successful in recognizing details than male students can be a source of DIF”.

Similarly, Expert 5 (E5) said: "I think that the question does not include item bias that causes female students to be advantageous… Possible sources that may cause DIF are not effective in this question. The act of writing a letter is not closer to female students or further away from male students in terms of cognitive, cultural, curriculum content, or socio-economic terms. In the image given regarding the question, a situation that provides an advantage to female students was not evaluated."

When the opinions of the other experts were examined, it was found that they made statements similar to those of experts 4 and 5. Considering that Expert 2 also gave an undecided opinion, it was found that Item 11 was not biased.

4. DISCUSSION and CONCLUSION

This study aimed to detect biased items in the TPSEE which was administered in the English subtest in April 2017. Since methods based on IRT were considered first, unidimensionality and local independence from IRT assumptions were tested. After providing the assumptions, the data fit of the model was examined according to 1PL, 2PL, and 3PL. While the best fit was found to be 3PL by all criteria, the first step was to determine if DIF with and without item purification was performed and when it was not performed based on MH, SIBTEST, Lord's χ^2, and Raju's unsigned area methods. In the analysis of IRT-based methods, DIF analyses were performed based on the 3PL model. When item purification was not performed based on at least two methods, only one item (item 11) was found to contain moderate (B level) DIF. When item purification was performed, DIF was detected at C level based on SIBTEST method and at B level according to MH in item 11. In the conceptual process-based bias analysis, the opinions of seven experts were sought on item 11 and six experts stated that the item was not biased. The reason for the DIF in item 11, where at least moderate DIF was found in three conditions, was that the females were more knowledgeable and had more vocabulary than the experts who were consulted. This situation indicated real differences, not bias.

When conducting DIF analysis using the four methods, it was found that the overall results were generally consistent but somewhat divergent. DIF analysis results varied with or without item purification. In the absence of item purification, DIF was detected in the same number of items based on the SIBTEST and MH methods. The MH and SIBTEST methods produced partially similar results. Compared to the Lord's chi-square method and the Raju's unsigned area test methods, DIF was detected in fewer items. When item purification was performed, the total number of DIF-containing items determined based on the four methods decreased. Except for item 11 all the items were unbiased items. One expert identified bias in item 11 but stated that she was not sure when deciding that during the face-to-face interviews.

The research findings indicate that the research results may differ based on the method used. Examination of DIF studies in the literature shows that the results may differ based on the methods (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019; Çepni & Kelecioğlu, 2021; Soysal &

Yılmaz Koğar, 2021). Camilli and Shepard (1994) suggest that DIF detection should be based on at least two methods. Thus, more accurate inferences can be made by comparing the results obtained from methods with different statistical backgrounds.

While there were studies in the literature that reveal the effect of purification on DIF detection (e.g., Özdemir, 2015; Tunc et al., 2018; Soysal & Yılmaz Koğar, 2021), it was seen that no bias study was conducted in any of these studies. Therefore, further bias studies are required. In his research on DIF detection using Lord's Chi-Square, Raju's Area and Likelihood-Ratio Test methods in the 2011 TIMSS mathematics subtest, Özdemir (2015) found that performing purification caused a difference in the number of items in which DIF was detected, especially in Lord's Chi-square and Raju's Area methods. When the item was purified, the number of DIF-detected items in these two methods decreased in this study, which is consistent with our study. However, purification or non-purification in the Likelihood-Ratio Test method did not cause such a difference. Soysal and Yılmaz Koğar (2021) found that DIF was detected in TPSEE 2016 Turkish subtest using Raju's unsigned area and Lord's χ^2 methods, and DIF was detected in more items when item purification was performed. Similarly, Tunc et al. (2018) reported the same findings as Soysal and Yılmaz Koğar's (2021) research; however, their results differ from those of our research.

When the studies in the literature were analyzed, it was seen that while there were numerous DIF detection studies, a limited number of studies on bias had been conducted (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018; Akcan & Atalay Kabasakal, 2019). Upon analysis of all these aforementioned studies, it was revealed that a bias study had been conducted for other courses' tests other than the English language test (e.g., Bakan Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018). In the literature, it was seen that the study of DIF and bias regarding the English language test was quite limited (e.g., Akcan & Kabasakal, 2019). Akcan and Kabasakal (2019) analyzed the items of the English test items of the "Undergraduate Placement Exam (UPE)" administered in 2016 by gender based on "MH, SIBTEST, and Multiple Indicator and Multiple Causes (MIMIC) methods". Their analysis of 60 items revealed that one item in the translation subtest was found to be DIF in favor of male students. Based upon expert opinion, they concluded that this item did not show bias. Using different DIF detection methods led to partially different conclusions in terms of the number of items with DIF and the level of DIF.

The research has several limitations and suggestions. Firstly, there were four methods utilized in the research. Therefore, similar studies could be performed using other DIF detection methods. TPSEE was administered in April 2017 and the "booklet A" dataset was used in the research. Similar studies can be performed on diverse datasets. As binary (1-0) data were used in the research, DIF detection can be conducted in polytomous data. Only gender-based bias has been addressed in recent research. Future researchers can focus on different sources of DIF. The study consulted the opinions of seven experts when determining the biased item. Similar studies can be designed by holding an item bias expert panel with the Delphi Technique. In the study, IRT-based DIF determination based on the 3PL model was performed. The results can be compared by making IRT-based DIF estimations based on 1PL, 2PL, and 3PL. No correction method was applied in this research. The effect of different correction methods on DIF detection can be investigated. This research was carried out based on real data, while simulation studies can be conducted under different conditions.

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.

Orcid

Serap Büyükkıdık https://orcid.org/0000-0003-4335-2949

REFERENCES

Akcan, R., & Atalay Kabasakal, K.A. (2019). An investigation of item bias of English test: The case of 2016 year undergraduate placement exam in Turkey. *International Journal of Assessment Tools in Education*, *6*(1), 48-62. https://doi.org/10.21449/ijate.508581

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

Bakan Kalaycıoğlu, D. (2022). Gender-based differential item functioning analysis of the medical specialization education entrance examination. *Journal of Measurement and Evaluation in Education and Psychology*, *13*(1), 1-13. https://doi.org/10.21031/epod.998592

Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Item bias analysis of the university entrance examination. *Education and Science*, *36*(161), 3–13.

Camilli, G. & Shepard, A.L. (1994). *Methods for identifying biased test items* (1st ed.). Sage.

Chalmers, R.P. (2018). Improving the crossing-SIBTEST statistic for detecting non-uniform DIF. *Psychometrika*, *83*(2), 376-386.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory* (1st ed.). Holt, Rinehart and Winston.

Çepni, Z., & Kelecioğlu, H. (2021). Detecting differential item functioning using SIBTEST, MH, LR and IRT methods. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(3), 267-285. https://doi.org/10.21031/epod.988879

Emily, D., Brooks, G., & Johanson, G. (2021). Detecting differential ıtem functioning: Item response theory methods versus the Mantel-Haenszel procedure. *International Journal of Assessment Tools in Education*, *8*(2), 376-393. https://doi.org/10.21449/ijate.730141

Fidalgo, A.M., Mellenbergh, G.J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43-53.

Freelon, D. (2013). ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service. *International Journal of Internet Science*, *8*(1), 10-16.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313-334. https://doi.org/10.1207/s15324818ame0204_4

Holland, P.W., & Thayer, D.T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, (2), i-24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Holland, P.W., & Wainer, H. (Eds.) (1993). *Differential item functioning* (1st ed.). Lawrence Erlbaum.

Karakaya, İ. (2012). An investigation of item bias in science and technology subtests and mathematic subtests in Level Determination Exam. *Educational Sciences: Theory and Practice*, *12*(1), 215–229.

Karakaya, İ., & Kutlu, Ö. (2012). An investigation of item bias in Turkish subtests in Level Determination Exam. *Education and Science*, *37*(165), 348–362.

Khalid, M.N., & Glas, C.A. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186-197. https://doi.org/10.1016/j.measurement.2013.12.019

Li, H.H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677

Llach, M.P.A., & Gallego, M.T. (2012). Vocabulary knowledge development and gender differences in a second language. *Elia*, *12*(1), 45-75.

Lord, F.M. (1980). *Applications of item response theory to practical problems* (1st edition). Erlbaum.

Magis, D., & Facon, B. (2013). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement*, *73*(2), 293-311. https://doi.org/10.1177/0013164412451903

Martinkova, P., & Drabinova, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503-515. https://doi.org/10.32614/RJ-2018-074

Osterlind, S.J. (1983). *Test item bias* (1st ed.). Sage.

Özdemir, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences*, *174*, 2075-2083. https://doi.org/10.1016/j.sbspro.2015.02.004

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495-502. https://doi.org/10.1007/BF02294403

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207. https://doi.org/10.1177/014662169001400208

Roussos, L., & Stout, W. (1996) A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*, 355-371. https://doi.org/10.1177/014662169602000404

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194.

Sireci, S.G., & Rios, J.A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187. https://doi.org/10.1080/13803611.2013.767621

Soysal, S., & Yılmaz Koğar, E.Y. (2021). An investigation of item position effects by means of IRT-based differential item functioning methods. *International Journal of Assessment Tools in Education*, *8*(2), 239-256. https://doi.org/10.21449/ijate.779963

Tunc, E.B., Uluman, M., & Avcu, A. (2018). Revisiting the effect of ıtem purification on differantial ıtem functioning; real data findings. *International Online Journal of Educational Sciences*, *10*(5), 139- 147. https://doi.org/10.15345/iojes.2018.05.010

Wang, W.C., & Su, Y.H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113-144. https://doi.org/10.1207/s15324818ame1702_2

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods* [Dissertation, Umea University]. Umea University Libraries EM No 60.

Yıldırım, H., & Büyüköztürk, Ş. (2018). Using the delphi technique and focus-group interviews to determine item bias on the mathematics section of the Level Determination Exam for 2012. *Educational Sciences: Theory & Practice*, *18*(2), 447-470.

Zieky, M. (1993). *Practical questions in the use of DIF statistics in test development*. In P. W. Holland, & H. Wainer, Differential Item Functioning (pp. 337-347). Erlbaum.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters, 160. https://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 781–795

https://doi.org/10.21449/ijate.1368341

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Type I error and power rates: A comparative analysis of techniques in differential item functioning

Ayse Bilicioglu Gunes[1*], **Bayram Bicak**[2]

[1]TED University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[2]Akdeniz University, Faculty of Education, Department of Measurement and Assessment, Antalya, Türkiye

ARTICLE HISTORY

Received: Sep. 29, 2023
Revised: Dec. 02, 2023
Accepted: Dec. 09, 2023

Keywords:
Classical test theory,
Item response theory,
Differential item functioning,
Type I error,
Statistical power.

Abstract: The main purpose of this study is to examine the Type I error and statistical power ratios of Differential Item Functioning (DIF) techniques based on different theories under different conditions. For this purpose, a simulation study was conducted by using Mantel-Haenszel (MH), Logistic Regression (LR), Lord's χ^2, and Raju's Areas Measures techniques. In the simulation-based research model, the two-parameter item response model, group's ability distribution, and DIF type were the fixed conditions while sample size (1800, 3000), rates of sample size (0.50, 1), test length (20, 80) and DIF- containing item rate (0, 0.05, 0.10) were manipulated conditions. The total number of conditions is 24 (2x2x2x3), and statistical analysis was performed in the R software. The current study found that the Type I error rates in all conditions were higher than the nominal error level. It was also demonstrated that MH had the highest error rate while Raju's Areas Measures had the lowest error rate. Also, MH produced the highest statistical power rates. The analysis of the findings of Type 1 error and statistical power rates illustrated that techniques based on both of the theories performed better in the 1800 sample size. Furthermore, the increase in the sample size affected techniques based on CTT rather than IRT. Also, the findings demonstrated that the techniques' Type 1 error rates were lower while their statistical power rates were higher under conditions where the test length was 80, and the sample sizes were not equal.

1. INTRODUCTION

Measurement tools define the levels of traits or qualities that individuals possess. Therefore, the measures obtained from them must be accurate and precise. Two fundamental properties must be present in a measurement tool: reliability and validity. Reliability refers to the stability and consistency of measurements. On the other hand, validity is a matter of whether the instrument can measure the intended characteristic. Bias is one of the threats to validity (Clauser & Mazor, 1998; Zumbo, 1999). In the administration process of some tests, measurement bias may occur due to factors such as the characteristics of the participants or culture. The responses of individuals with the same ability level but in different subgroups to an item or a test may differ. This often results in the item or test functioning differently for people in different subgroups (Dorans & Holland, 1993). Exposure to item bias has been shown to have adverse effects on validity. Camilli and Shepard (1994) define bias as the systematic error in test scores

*CONTACT: Ayse Bilicioglu Gunes ✉ ayse.bilicioglu@tedu.edu.tr TED University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

toward a certain group. If this situation, which gives an advantage to one group and a disadvantage to another group at the same ability level, occurs in only a few items of the test, it is called item bias, and if it is a circumstance that occurs throughout the test, it is called test bias (Zumbo, 1999). In item bias, the probability of responding correctly to an item depends on belonging to a group rather than the level of ability that is being measured. This is one of the main signs of measurement error (Osterlind, 1983). If test scores do not reflect the intended construct in the same way between groups or situations (or indicate different constructs for different groups), score interpretations will be biased.

The first stage of bias studies is to determine differential item functioning (DIF) as an index of bias (Camilli & Shepard, 1994). According to Hambleton et al. (1991), DIF is the difference in the probability of responding correctly to an item when individuals in different groups are at the same ability level. Two groups -the focus group and the reference group- can be examined for determining DIF analysis. In the focus group, while answering the item, the unfavorable circumstances of people with similar abilities are studied. The reference group is the group that the focus group is compared to (Zumbo, 1999).

Camilli and Shepard (1994) note two potential reasons for the incidence of DIF. These are called item effect and item bias. Item effect is the actual difference between the probabilities of respondents in different groups giving a correct answer to an item (Zumbo, 1999). This difference can be attributed to the prior experience or knowledge of one of the groups. If the participants differ in terms of the knowledge they hold, it is expected that the responses to items will also reflect this difference. DIF is a necessary, but not sufficient, condition for item bias. Thus, there is no item bias if DIF is not apparent for a given item. But even if DIF is obvious, its existence alone does not prove item bias; additional item bias analyses (such as content analysis and empirical evaluation) are required to establish the existence of item bias (Zumbo, 1999). In item bias, factors other than the ability to be measured are included in the test during the measurement. The purpose of determining DIF is to explain the differences originating from this bias (Dorans & Holland, 1993). DIF occurs in two forms: uniform DIF and non-uniform DIF (Mellenbergh, 1983). When groups and ability levels do not interact, uniform DIF appears. Conversely, an interaction is involved in non-uniform DIF (Swaminathan & Rogers, 1990).

Many techniques are used to determine uniform and non-uniform DIF. The techniques used vary depending on whether the data are dichotomous or polytomous. Techniques are broadly categorized under two theories. While DIF determination techniques based on Classical Test Theory (CTT) compare the distribution of the groups' scores, those based on Item Response Theory (IRT) compare the probability of correct response of the groups to the related item. Among the frequently used DIF determination techniques based on CTT are Mantel-Haenszel (MH), Analysis of Variance, Transformed Item Difficulty, and Logistic Regression (LR). Also, some of the frequently used techniques based on IRT are the Likelihood Ratio Test, Raju's Area Measures, and Lord's χ^2 (Camilli & Shepard, 1994; Hambleton et al., 1991).

The MH technique was developed by Nathan Mantel and William Haenszel in the 1950s as a chi-square method. Afterwards, it was updated by Holland and Thayer (1998) and introduced to identify DIF. The MH technique is a frequently utilized technique as it does not require large samples, can provide effect size values, and calculations are not complex (Samuelsen, 2005). The values obtained with this technique are interpreted with the delta scale (Δ) taking into account the category levels recommended by The Educational Test Service (ETS). LR, which is another method based on CTT, is an expanded version of the MH technique and a sensitive technique for identifying both uniform and non-uniform DIF. The technique is based on a standard logistic regression model that uses independent variables to predict two dependent variables. LR assesses the presence of DIF by utilizing responses to the item and the total score. LR is a widely used technique due to its simplicity of programming and robustness in dealing with non-uniform DIF (Clauser & Mazor, 1998; Swaminathan & Rogers, 1990). Raju's Area Measures is a technique that evolved based on IRT. This technique takes into account the Item Characteristic Curves (ICCs) in the determination of DIF (Raju, 1988). The area mentioned in Raju's Area Measures is the gap between the estimated ICCs of two groups (Camilli & Shepard, 1994). In Lord's χ^2 method, the differences in item parameters of two different groups for an item are compared (Lord, 2012). If there is a discrepancy between the two groups as a result of the comparison, it can be said that the relevant item functions in a different way. One of the virtues of the technique is that it can be used to determine both uniform and non-uniform DIF (Price, 2014).

As pointed out by Crocker and Algina (1986) and Dorans and Holland (1993), if the items in the test are biased, it means that the decisions based on these test scores also contain errors. Therefore, it is very pivotal to elaborate on how the techniques used to determine bias work and under which conditions. Questions such as "What is the main reason for an item to include DIF?" and "What should be done if an item contains DIF?" embody the essential questions of DIF studies. Researchers have discussed these questions and offered solutions like removing the relevant item from the test and adding a new one. However, this process is quite time-consuming and also affects the content validity of the test. Another argument is that the item showing DIF should be revised. This suggestion requires the researchers to implement the revised items to a group of students and then conduct a DIF analysis again (Ellis & Raju, 2003). Given these points, studies to detect the sources of DIF and mitigate the presence of item bias are imperative. Experts' opinion is the most widely used technique to investigate whether the item is biased or not as a consequence of DIF analyses. However, in some occasions, although the item is not biased, results can give the opposite information, and at that time the question appears “Why does the item have DIF?”. One possible reason for this may be Type I error rates. The presence of Type I error can be considered as a misidentification of DIF in the scope of item bias. In this case, even though the item does not actually possess DIF, it gives a statistically significant output as containing DIF. In other words, although the item displays similar performance across individuals in the focal and reference groups, the technique used indicates that the item displays DIF (Dainis, 2008). This problem has been the subject of research for various reasons (Jodoin & Gierl, 2010). The first reason is that identifying an item as presenting DIF is a waste of time and resources. If the DIF identification is faulty, it wastes resources in the test development process. Secondly, researchers waste their time on DIF analyses. Ultimately, from the perspective of the research, if the study intends to determine the strength of DIF detection techniques or to identify DIF items from real data, inferences drawn from comparisons (in terms of the quality and effectiveness of the technique) are not be valid due to Type I errors.

Studies on the effectiveness of DIF detection techniques taking certain variables into account have been released into the literature. Ankenmann et al. (1999) aimed to compare the Type I error and statistical power rates of MH and LR techniques. The results indicated that the MH technique had better statistical power compared to the LR technique, but both techniques were influenced by sample sizes. In terms of Type I error rates, it was found that the MH technique exceeded the nominal alpha level. In a research conducted by Gierl et al. (2000), the performance of MH, LR, and SIBTEST techniques was compared under different conditions such as the presence of DIF, sample size, and ability distribution. The results suggested that even in cases with small sample sizes, the Type I error rates of all three techniques were around the nominal alpha level. The SIBTEST technique, however, exhibited the highest statistical power. Dainis (2008) compared methods based on the CTT and IRT in terms of Type I error and power rates. The study found that the LR technique yielded low power and high Type I error rates. In a study by Demars (2009) comparing the Type I error rates of MH, LR, and SIBTEST techniques under different conditions, it was found that reducing test length and increasing sample size led to inflated Type I error rates for MH and LR techniques. Vaughn and Wang (2010) investigated the Type I error rates of classification trees, MH, and LR techniques under conditions of sample size, DIF, and ability distribution. The study concluded that MH and LR techniques had low Type I error rates for three different sample sizes and ratios. Magis and De Boeck (2012) examined the performance of the MH technique under different conditions and found that the MH technique yielded inflated Type I error rates in situations where there was a between-group ability difference and when sample size increased. Atalay Kabasakal et al. (2014) compared the Type I error rates and power of MH, SIBTEST, and MTK-OO techniques under different conditions. The results indicated that in all conditions considered, the MH technique had the highest power, while the SIBTEST technique had the highest error. In a study by Sunbul and Omur Sunbul (2016), MH, LR, Lord's χ^2, and Raju's area measures were compared in terms of Type I error and power. The results suggested that techniques based on CTT were not significantly affected by varying conditions in Type I error rates, and both theory-based techniques showed an increase in power with an increase in sample size.

When the studies were examined, it was seen that a number of research has been published on the performance of DIF techniques considering various conditions, but most of it has been conducted with techniques based solely on the CTT or IRT (Ankenmann et al., 1999; Cohen et al., 1996; Demars, 2009; Gierl et al., 2000; Jodoin & Gierl, 2010; Kristjansson et al., 2005; Lim & Drasgow, 1990; Magis & De Boeck, 2012; Vaughn & Wang, 2010). Few published studies have used both (Atalay Kabasakal et al., 2014; Atar, 2007; Dainis, 2008; Erdem Keklik, 2012; Finch & French, 2007). Although MH, LR, Raju's Area Measures, and Lord's χ^2 techniques are frequently utilized in the literature, to date, there has been little comparative research conducted on Type I errors and powers of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques at once (Basman, 2023; Sunbul & Omur Sunbul, 2016). In addition, since the presence of Type I error can be considered as misidentification of DIF within the scope of item bias and statistical power shows the performance of the techniques, this study aimed to investigate the results of MH, LR, Raju's Area Measures and Lord's χ^2 techniques under different conditions and to compare the techniques with each other by considering Type I error and power ratios during the comparison of the techniques.

In this scope, it is thought that the study's results will help determine the appropriate DIF determination techniques that can be used for institutions working with large-scale tests, test developers, and those who make decisions based on the scores of the relevant tests.

1.1. Purpose of the Study

The main purpose of this study is to examine the Type I error and statistical power ratios of DIF determination techniques based on different theories under different conditions for items scored with two categories. In pursuit of this objective, the study seeks to address the following questions:

1. How do the Type I error rates of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under conditions where the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and there are no DIF items?
2. How do the statistical power ratios of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under the conditions in which the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and the ratio of DIF- containing items is 0.05 and 0.1?

2. METHOD

Within the scope of the research, Type I error and statistical power ratios were scrutinized by conducting DIF analyses with the data generated in the 2PL model under different conditions. The research follows a simulation-based model that provides the researcher with the opportunity to work under different conditions (Dooley, 2002).

2.1. Generation of Data

The conditions held constant in the current study are the IRT model used, the ability distributions of the focus and reference groups, and the type of DIF. The conditions manipulated include sample size and proportions, test length, and the proportion of items containing DIF.

2.1.1. *Fixed conditions*

In this study, participant responses, ability levels, and item parameters for the focus and reference groups were generated in compliance with the 2PL model. Since the data fit of the 2PL model was better than that of the 1PL model, the 1PL model was not included in this study. When the 3PL model was examined under real test conditions, it was noted that the standard error in the estimations increased due to the c parameter, and therefore, it is not a strong statistic in DIF determination studies (Hambleton et al., 1991). Considering these reasons, only the 2PL model was included in this study. The ability distribution of the groups was fixed using a normal distribution with a mean of 0 and a standard deviation of 1 (Dodeen, 2004; Hauck Filho et al., 2014; Roussos & Stout, 1996).

In generating the parameters, they were held constant for both groups, while the b parameter of the reference group was altered. Data were generated using R software. For both groups, the "a" parameter was obtained from a normal distribution with a mean of 0.8 and a standard deviation of 0.04 (Sunbul & Omur Sunbul, 2016), while the "b" parameter was randomly drawn from a distribution with minimum and maximum values of -2 and +2, respectively (Desa, 2012; Kogar, 2018). In studies conducted on simulated data in the literature, various values for the amount of DIF have been utilized, such as 0.10, 0.15, 0.25, 0.30, 0.32, 0.43, 0.53 (Atar, 2007; Fidalgo et al., 2000; Kristjansson, 2001; Wang & Su, 2004; Zwick et al.,1993). In the current study, production of DIF-containing items was conducted by adding 0.05 and 0.10 values to the b parameter. Uniform DIF was obtained in this way, and the research was carried out.

2.1.2. *Manipulated conditions*

Swaminathan and Rogers (1990) stated that one of the factors that may affect statistical estimation in studies is the sample size and sample size ratios of the focus and reference groups. Additionally, it has been noted that nonparametric techniques have enhanced power to identify items with DIF when the sample sizes of the groups are not equal (Kristjansson et al., 2005). When reviewing the literature, it is observed that techniques based on CTT have sufficient power when there are at least 200-250 individuals per group (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993). In techniques based on IRT, at least 1000 samples are expected (Shepard et al., 1981). Similarly, according to the results obtained from their study, Gök and Kelecioğlu (2014) state that more stable estimates can be obtained as the sample size increases, and sample sizes of 1000 and above would be sufficient. Additionally, considering large-scale testing, it is known that very large sample sizes are used. Considering these factors and this study examined techniques belonging to two different theories, and one of these theories is IRT, which requires large samples, sample sizes of 1800 and 3000 were chosen in the study. The ratio of the sample size of the reference group to the total sample size was manipulated as R/T1= 0.50 and R/T2= 0.75. Another parameter analyzed within the scope of DIF is test length. In other studies in the literature, it is seen that the number of items is generally set as 20, 40, and 80 (Atalay Kabasakal et al., 2014; Narayanan & Swaminathan, 1994; Price, 2014; Wang et al., 2013). Test lengths of 20 can be considered as small, 40 as medium, and 80 as considerable. Regarding national exams in Turkey, it is known that Liselere Giriş Sınavı (LGS) consists of 10-20 items, while exams such as Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES), Yabancı Dil Sınavı (YDS), and Yükseköğretim Kurumları Yabancı Dil Sınavı (YOKDIL) consist of 80-100 items. Given the large-scale exams in Turkey, two different test lengths, 20 and 80, were utilized in this study. The selection of the proportion of items containing DIF was based on the rates of 0, 0.05 and 0.10, as utilized in the studies by Atalay Kabasakal et al. (2012) and Sunbul and Omur Sunbul (2016).

Within the scope of this research, data were generated with 25 replications for a total of (2x2x2x3) 24 conditions, including two sample sizes, two sample size ratios, two test lengths, and three ratios of DIF-containing items, since it is known that errors decrease after 25 replications in data generation (Harwell et al., 1996).

2.2. Data Analysis

In the current study, data were generated and analysed using R software. In this study, LR and MH techniques based on CTT and Lord's χ^2 and Raju's Area Measures based on IRT were used for dichotomous data. Statistical analysis was performed using the "difR" package in the R software, which includes DIF detection techniques and facilitates the analysis of these techniques both independently and comparatively (Magis et al., 2018). The "ltm" package, which also provides model estimation, was used for the analysis of the techniques based on IRT (Rizopoulos, 2022).

DIF detection techniques were assessed with two criteria: Type I error and statistical power ratios. Following the analyses, for Type I error, the proportion of items that did not exhibit DIF but showed an analysis output indicating DIF was calculated and reported. In terms of statistical power ratios, the percentage of items that actually contained DIF, whereas the result of the analysis indicated the presence of DIF, was reported. For the Type I error rates of the results obtained from the techniques used, the classification in Table 1 by Bradley (1978) was considered. In terms of statistical power ratios, the criterion stipulated that techniques displaying values exceeding 0.80 were deemed sufficient and high, whereas those falling below this threshold were considered inadequate (Atalay Kabasakal et al., 2014).

Table 1. *Bradley's classification of type I error rates.*

Level	Value Range
Conservative	$\alpha < 0.045$
Maintained	$0.045 < \alpha < 0.055$
Inflated	$\alpha > 0.055$

3. FINDINGS

The findings related to the first sub-problem " How do the Type I error rates of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under conditions where the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and there are no DIF items?" are presented in Table 2.

Table 2. *Findings related to the first sub-problem.*

Conditions			DIF Techniques				
The ratio of Items Showing DIF	Sample Size	Number of Items	R/ T	MH	LR	LORD χ^2	RAJU
0 (There is no DIF)	1800	20	0.50	0.92	0.92	0.86	0.74
			0.75	0.90	0.89	0.79	0.68
		80	0.50	0.89	0.88	0.88	0.79
			0.75	0.86	0.86	0.86	0.79
	3000	20	0.50	0.92	0.92	0.89	0.78
			0.75	0.89	0.89	0.86	0.78
		80	0.50	0.91	0.91	0.90	0.84
			0.75	0.91	0.91	0.91	0.85

It can be seen from the data in Table 2 that the Type I error ratios under different conditions for 20 and 80 items with a sample size of 1800 vary between 0.68 and 0.92 and between 0.79 and 0.89, respectively. For 20 items, when sample size ratios are considered, the MH technique yielded the highest error, and Raju's Area Measures technique yielded the lowest error. This pattern was consistent for 80 items and a sample size ratio of 0.50. When the sample size ratio was set as 0.75, the Lord’s χ^2 technique showed the highest error with 0.862, while Raju's Area Measures had the lowest error with α value of 0.791.

When scrutinizing the Type I error rates for 20 and 80 items under diverse conditions, with sample sizes of 3000, values ranged from 0.78 to 0.92 and 0.84 to 0.91, respectively. Notably, in both conditions involving 20 and 80 items, it is observable that the MH and LR methods consistently exhibited the highest error rates, whereas Raju's Area Measures technique consistently manifested the lowest error rates, particularly when sample size ratios were altered. However, upon closer analysis, it is evident that the error rates of techniques other than Raju's Area Measures are very close to each other.

The findings related to the second sub-problem " How do the statistical power ratios of MH, LR, Raju's Area Measures, and Lord's χ^2 techniques differ under the conditions in which the sample size is 1800 and 3000, the sample size ratio is 0.50 and 0.75, the number of items is 20 and 80, and the ratio of DIF- containing items is 0.05 and 0.1?” are presented in Table 3.

Table 3. *Findings related to the second sub-problem.*

Conditions			DIF Techniques				
The Ratio of Items Showing DIF	Sample Size	Number of Items	R/ T	MH	LR	LORD χ^2	RAJU
0.05	1800	20	0.50	0.80	0.68	0.64	0.52
			0.75	0.64	0.44	0.56	0.44
		80	0.50	0.11	0.12	0.41	0.39
			0.75	0.73	0.71	0.68	0.70
	3000	20	0.50	0.12	0.08	0.28	0.20
			0.75	0.20	0.12	0.12	0.04
		80	0.50	0.08	0.12	0.32	0.28
			0.75	0.85	0.78	0.60	0.58
0.1	1800	20	0.50	0.60	0.54	0.50	0.38
			0.75	0.20	0.14	0.34	0.32
		80	0.50	0.09	0.10	0.31	0.24
			0.75	0.59	0.51	0.53	0.51
	3000	20	0.50	0.14	0.08	0.18	0.20
			0.75	0.08	0.04	0.42	0.38
		80	0.50	0.77	0.76	0.60	0.53
			0.75	0.13	0.14	0.48	0.44

As shown in Table 3, the statistical power ratios ranged between 0.14 and 0.80 for a sample size of 1800 and 20 items under varying conditions, including sample size and the proportion of items with DIF. When the proportion of items containing DIF was set at 0.05, it became evident that the MH method had the highest power ratios, while Raju's Area Measures technique had the lowest power. However, under conditions where the DIF item proportion increased to 0.1, and sample size ratios stood at 0.50 and 0.75, MH and Lord's χ^2 exhibited the highest power with 0.60 and 0.34, respectively. Conversely, when considering sample size ratios of 0.50 and 0.75, Raju's Area Measures and the LR technique displayed the least powerful performance.

The statistical power ratios for a sample size of 1800 and 80 items ranged between 0.11 and 0.73 under varying conditions of sample size and DIF-containing item ratios. When the proportion of DIF-containing items was 0.05 and the sample size ratios ranged between 0.50 and 0.75, Lord's χ^2 and MH techniques had the highest power, while MH and Lord's χ^2 techniques had the lowest power, respectively. For 80 items, when the proportion of items containing DIF increased to 0.1, and the sample size ratios were 0.50 and 0.75, Lord's χ^2 with 0.31 and MH with 0.59 were found to have the highest power. In addition, when the sample size ratio was 0.50, MH with 0.09, and when the sample size ratio was 0.75, LR and Raju's Area Measures with 0.51 were found to have the lowest power.

For a sample size of 3000 and 20 items, statistical power ratios ranged between 0.04 and 0.42 under varying conditions of sample size and DIF-containing item ratios. Raju's Area Measures technique and MH technique were found to have the highest power ratios when the DIF-containing item ratio was 0.05 and the sample size ratios were 0.50 and 0.75, respectively. For the 0.50 sample size ratio, the LR technique with 0.08 and for a 0.75 sample size ratio, Raju's Area Measures technique with 0.04 were found to have the lowest power. For 20 items, Raju's Area Measures and Lord's χ^2 techniques were found to have the highest power when the DIF-containing item ratio was 0.1 and the sample size ratios were 0.50 and 0.75, respectively, while the LR technique was found to have the lowest power with 0.08 for a sample size ratio of 0.50.

For a sample size of 3000 and 80 items, statistical power ratios ranged between 0.08 and 0.85 under varying conditions of sample size and the ratio of items showing DIF. Lord's χ^2 and MH techniques had the highest power when the proportion of items with DIF was 0.05 and the sample size ratios were 0.50 and 0.75, respectively. MH technique had the lowest power with 0.08 for a sample size ratio of 0.50, and Raju's Area Measures technique had the lowest power with 0.58 for a sample size ratio of 0.75. When the proportion of items containing DIF was 0.1 and the sample size ratios were 0.50 and 0.75, the MH technique and Lord's χ^2 had the highest power, respectively, while Raju's Area Measures had the lowest power with 0.53 and MH with 0.13 when the sample size ratios were 0.50.

4. DISCUSSION and CONCLUSION

Within the scope of the current research, Type I error and statistical power ratios were scrutinized by conducting DIF analyses with the data generated in the 2PL model under different conditions. As a result of the analyses performed for this purpose, it was found that Type I error rates were higher than the nominal error level (0.05) in all conditions. Based on these results, it can be inferred that there were inflated Type I errors. For the 1800 sample size, when all conditions are analyzed, the MH technique displayed the highest Type I error rates. When the techniques used were assessed within the conditions, it was found that the MH and LR techniques produced very similar values and exhibited higher errors in the condition where the number of items was 20 compared to the condition where the number of items was 80 for a sample size of 1800. Both techniques yielded the highest error rate for the condition where the number of items was 20 and the R/T ratio was 0.50, and the lowest error rate for the condition where the number of items was 80 and the R/T ratio was 0.75. These results align with the findings of other studies. Kim (2010) studied the Type I error rates of LR, MH, DFIT, and Lord's χ^2 techniques under different conditions, and found that all techniques tended to inflate Type I errors in conditions where the test length was shorter, the sample size was larger, and the focal and reference groups were equal. Similarly, Demars (2009) observed that MH and LR techniques produced inflated Type I error rates with shorter test lengths and larger sample sizes. On the other hand, in a study conducted by Dainis (2008), a comparison of the Type I errors of the IRT-OO, DFIT, MH, and LR techniques revealed that the error rates of the MH technique were at an acceptable nominal level. Gierl et al. (2000) emphasized that the Type I error rates of the MH and LR techniques can be around or even below the nominal error levels even if the sample size is small with equal ability distributions. Ankenmann et al. (1999) also stated that the LR technique yields Type I error rates at the nominal error level under general conditions, while the MH technique yields error rates above this nominal error level. However, in the current study, both techniques yielded Type I error rates much higher than acceptable error rates in all conditions where the sample size varied, and the findings of the current study do not support the previous research (Ankenmann et al., 1999; Dainis, 2008; Gierl et al., 2000). The main reason for this situation might be the influence of the sample size on the employed techniques. When examining studies in the literature, it has been observed, especially for MH and LR methods, that inflated Type I error values are obtained as the sample size increases (Demars, 2009; Kim, 2010; Magis & De Boeck; 2012). In the context of this study, considering the methods based on MTK were employed, sample sizes exceeding 1000 were chosen. It is believed that the high Type I error values in the obtained findings are attributable to this choice.

For a sample size of 3000, upon comprehensive examination of all conditions, it is evident that LR had the highest Type I error rates. However, it should be noted that the differences in error rates among the various techniques were modest, indicating that the MH technique also exhibited high Type I error rates. All the values obtained, however, showed an inflated Type I error, just as in the conditions generated with a sample size of 1800.

In the conditions set up to determine the Type I error rates, broadening the sample size from 1800 to 3000 increased the Type I error values in all conditions except three. The three mentioned conditions and techniques are as follows: For 20 items with an R/T ratio of 0.75, the MH technique yielded a lower Type I error rate. Again, when the R/T ratio was 0.50 for 20 items, the MH technique produced the same error rate value for 1800 and 3000 sample sizes, while the LR technique produced the same error rate value when the R/T ratio was 0.75. When the techniques used within the context of the research were analyzed within the conditions, MH and LR techniques obtained very similar values in all conditions, as in the 1800 sample size. Both techniques yielded the lowest Type I error rates in the condition where the number of items was 20, and the R/T ratio was 0.75. When the findings of Lord's χ^2 and Raju's Area Measures techniques were examined, it was concluded that both techniques yielded the highest Type I error rates under the condition where the number of items was 80, and the R/T ratio was 0.75. The lowest Type I error rates were obtained under the condition where the number of items was 20, and the R/T ratio was 0.75, just as in the techniques based on CTT.

Comparing the Type I error findings obtained based on the theories they are grounded in, it can be interpreted, similar to the study by Kan et al. (2013), that the techniques based on CTT and IRT are similar in terms of the conditions under which they are affected. Techniques in both theories display lower Type I error rates under the condition of 0.75, where the sample sizes of the focal and reference groups are not equal. Erdem Keklik (2012) stated that lower Type I error rates were obtained when the sample size ratios were 1:2 rather than 1:1. Comparison of the findings with those of other studies (Demars, 2009; Magis & DeBoeck, 2012) confirms that Type I error rates tend to increase with growing sample size; however, this does not appear to be the case for some studies (Dainis, 2008; Sunbul & Omur Sunbul, 2016; Vaughn & Wang, 2010) that report that error rates decrease.

In the context of the first sub-problem, it is plausible to attribute the elevated Type I error rates observed in the various conditions examined within the research to the specific circumstances established. These conditions can be seen as a key contributing factor to the deviation of Type I error rates from the nominal error level. These results are likely related to the sample size and test length used. The sample sizes used in the study were not small; large samples were employed, and the finding that large sample sizes increase Type I error rates is supported by other studies in the literature (Dainis, 2008; Demars, 2009; Gierl et al., 2000; Magis & De Boeck; 2012). Regarding test length, higher Type I error rates were obtained in the analyses with 20 items compared to the conditions with 80 items. It is possible that these results were influenced by the techniques based on CTT, yielding higher rates due to the theoretical structure of the related techniques. The first point to be mentioned is that the MH and LR techniques are based on the observed score. These techniques use true scores as the matching variable in the process of determining the DIF. As Zwick et al. (1997) stated in the evaluation of the techniques used in this respect, the measurement errors of the techniques based on observed scores may decrease the reliability of the test. Therefore, this leads to true scores that deviate from the mean. Another problem that can be referred to is the use of total test scores, working with observed score-based techniques such as MH and LR with data produced following IRT models may cause inflated Type I errors.

The findings related to the second sub-problem, statistical power ratios, revealed that the MH technique yielded the highest statistical power ratio value of 0.80 as a result of the analyses conducted for a sample size of 1800. The MH technique displayed the highest statistical power ratios in all conditions except for three conditions. Surprisingly, it reached its highest value when the number of items was 20, not 80 as expected. Similar to the results of the study, Atalay Kabasakal et al. (2014) also found that the MH technique had the highest statistical power ratios in all conditions. In another study, it was also confirmed that the MH technique was the most powerful technique (Kristjansson et al., 2005). Regarding the results based on the other techniques included in the study, LR, Lord's χ^2, and Raju's Area Measures techniques reached the highest statistical power ratios when the number of items was 80, the ratio of DIF-containing items was 0.05, and the R/T ratio was 0.75. However, the obtained ratios were distributed around 0.70 and could not reach the desired value of 0.80. The fundamental reason for this situation can be attributed to the sample size, similar to the Type I error rates. When examining studies in the literature, findings have been obtained indicating that the statistical power of DIF techniques is also negatively affected by the sample size (Ankenmann et al., 1999; Gierl et al., 2000; Atalay Kabasakal et al., 2014).

An overall evaluation of the techniques based on the theory they are grounded in reveals that, in five of the eight conditions within the scope of the second sub-problem, MH and LR techniques based on CTT were found to have higher values, while Lord's χ^2 and Raju's Area Measures techniques based on IRT were found to have higher values in three of them. From this perspective, since the techniques do not show a regular difference from condition to condition, based on the current study's results, it can be interpreted that the techniques based on CTT provide higher statistical power ratios at a sample size of 1800. Furthermore, a consistent trend is observed across all employed techniques, where alterations in the proportion of items containing DIF from 0.05 to 0.10, in conditions with both 20 and 80 items, negatively impact and reduce the statistical power ratios. However, in contrast to earlier findings, Atar and Kamata (2011) emphasized that the statistical power ratio decreases as the proportion of items containing DIF decreases, especially in the LR technique, if the sample size is small. Another study compared the performance of LR and MH techniques and found that the statistical power ratios increased as the DIF-containing item ratio increased (Hidalgo & Lopez-Pina, 2004). Therefore, a definitive interpretation regarding the DIF-containing item ratio cannot be made.

Following the analysis conducted with a sample size of 3000, wherein a comparison of statistical power ratios among the techniques was performed, it is noteworthy that the MH technique yielded the highest statistical power ratios, consistent with the findings observed for the 1800 sample size, with a statistical power ratio of 0.85. However, the technique displayed lower performance than the other techniques in most of the conditions compared to the 1800 sample size. At this point, it can be said that the MH technique is affected by the sample size (Ankenmann et al., 1999). In the literature, there are studies that further support the idea that MH and LR techniques decrease statistical power ratios with increasing sample size (Erdem Keklik, 2012; Vaughn & Wang, 2010). On the other hand, Atar (2007) and Jodoin and Gierl (2010) state that increasing sample size leads to an increase in both statistical power ratios and Type I error rates. This differs from the findings presented here.

As a whole, when considering the findings derived from the evaluation of techniques within the research across various conditions, it becomes evident that higher statistical power ratios were consistently achieved under nearly all conditions when the number of items was set to 80. Techniques obtained the highest statistical power ratios when the number of items was 80, the ratio of DIF-containing items was 0.05, and the R/T ratio was 0.75. In this respect, it can be interpreted that increasing the number of items has a promising effect on the statistical power ratios. Another point that can be mentioned here is the impact of the R/T ratio on the power ratios. Although higher statistical power ratios were obtained when the ratio was 0.75, as stated in the current study, it was also stated in different studies that there may be inconsistencies in the techniques in conditions where the sample sizes of the groups are not the same (Jodoin & Gierl 2010; Narayanan & Swaminathan, 1994). Therefore, a definite interpretation regarding the R/T ratio cannot be made.

In line with the comparison in terms of theories, it is apparent that, in general, higher statistical power rates were achieved at a sample size of 1800. This observation promotes the findings reported by Sunbul and Omur Sunbul (2016), where it was asserted that augmenting the sample size led to an increase in the statistical power ratios for both CTT and IRT-based techniques. In fact, there exist other studies, such as Atar (2007) and Narayanan and Swaminathan (1994), which concur with this notion, indicating that larger sample sizes tend to enhance power ratios. Nevertheless, it is notable that in the present study, the majority of techniques exhibited superior performance with a sample size of 1800, with only a few exceptions.

Based on the Type I error and power rates findings obtained in this study, researchers studying with smaller samples may consider using techniques based on CTT while those working with larger samples may prefer techniques based on IRT. Additionally, considering the R/F ratios used in this study, it was found that Type I error rates were lower under conditions where the sample sizes of the focal and reference groups were not equal. Therefore, if practitioners have the flexibility in forming groups, it is recommended to create groups with unequal sample sizes.

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.

Authorship Contribution Statement

Ayse Bilicioglu Gunes: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, Validation, and Writing-original draft. **Bayram Bicak**: Supervision.

Orcid

Ayse Bilicioglu Gunes https://orcid.org/0000-0002-1603-8631
Bayram Bicak https://orcid.org/0000-0003-0860-9374

REFERENCES

Ankenmann, R.D., Witt, E.A., & Dunbar, S.B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistics in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277–300. https://doi.org/10.1111/j.1745-3984.1999.tb00558.x

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Değişen madde fonksiyonunun belirlenmesinde MTK olabilirlik oranı, SIBTEST ve mantel- haenszel yöntemlerinin performanslarının (I. Tip hata ve güç) karşılaştırılması [Comparison of the performance (Type I error and power) of the IRT likelihood ratio, SIBTEST, and mantel-haenszel techniques in determining the differential item functioning]. *Educational Sciences: Theory & Practice*, *14*(6), 2175- 2193.

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* (FSU_migr_etd-0248) [Doctoral dissertation, Florida State University]. http://purl.flvc.org/fsu/fd/FSU_migr_etd-0248.

Atar, B., & Kamata, A. (2011). Comparison of IRT likelihood ratio test and logistic regression DIF detection procedures. *Hacettepe University Journal of Education*, (41), 36–47.

Basman, M. (2023). A comparison of the efficacies of differential item functioning detection methods. *International Journal of Assessment Tools in Education*, *10*(1), 145-159. https://doi.org/10.21449/ijate.1135368

Bradley, J.V. (1978). Robustness. *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. http://dx.doi.org/10.1111/j.2044-8317.1978.tb00581.x

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Sage Publications.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedure to identify differential item functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Cohen, A.S., Kim, S.H., & Wollack, J.A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. CBS College Publishing.

Dainis, A.M. (2008). *Methods for identifying differential item and test functioning: An investigation of type 1 error rates and power* (3323367) [Doctoral dissertation, James Madıson University]. ProQuest.

DeMars, C.E. (2009). Modification of the mantel-haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, *34* (2), 149- 170.

Desa, Z.N. (2012). *Bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification* (3523517) [Doctoral thesis, University of Kansas]. ProQuest.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement*, *41*(3), 261- 270.

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). Blackwell

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-haenszel and standardization. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Lawrence Erlbaum.

Ellis, B.B., & Raju, N.S. (2003). Test and item bias: What they are, what they aren't, and how to detect them. *Educational Resources Information Center (ERIC).*

Erdem Keklik, D. (2012). *İki kategorili maddelerde tek biçimli değişen madde fonksiyonu belirleme tekniklerinin karşılaştırılması: Bir simülasyon çalışması [Comparison of techniques in detecting uniform differential item functioning in dichotomous items: A simulation study]* (311744) [Doctoral thesis, Ankara University]. YÖK, Ulusal Tez Merkezi.

Fidalgo, A.M., Mellenberg, G.J., & Muniz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, *5*(3), 43–53.

Finch, W.H., & French, B.F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565-582. https://doi.org/10.1177/0013164406296975

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000). Performance of mantel-haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large. The Annual Meeting of the American Educational Research Association.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage Publications.

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125. https://doi.org/10.1177/0146621696020002

Hauck Filho, N., Machado, W.D.L., & Damásio, B.F. (2014). Effects of statistical models and items difficulties on making trait-level inferences: A simulation study. *Psicologia: Reflexão e Crítica*, *27*(4), 670- 678. https://doi.org/10.1590/1678-7153.201427407

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. https://doi.org/10.1177/0013164403261769

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Ed.), *Test validity* (pp.129-145). Erlbaum.

Jodoin, M.G., & Gierl, M.J. (2010). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Psychological Measurement*, *14* (4), 329- 349. https://doi.org/10.1207/S15324818AME1404_2

Kan, A., Sünbül, Ö., & Ömür, S. (2013). 6. - 8. sınıf seviye belirleme sınavları alt testlerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi [Analysis of 6th - 8th grade placement exams subtests' differential item functioning by various methods]. *Mersin University Journal of the Faculty of Education*, *9*(2), 207- 222.

Karasar, N. (2010). *Bilimsel araştırma yöntemleri [Research methods]*. Nobel Publication.

Kim, J. (2010). *Controlling type I error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing* [Doctoral thesis, Georgia State University]. https://doi.org/10.57709/1642363

Koğar, H. (2018). An examination of parametric and nonparametric dimensionality assessment methods with exploratory and confirmatory models. *Journal of Education and Learning*, *7*(3), 148-158. 10.5539/jel.v7n3p148

Kristjansson, E. (2001). *Detecting DIF in polytomous items: an empirical comparison of the ordinal logistic regression, logistic discriminant function analysis, Mantel, and* *generalized Mantel Haenszel procedures* [Unpublished Doctoral Dissertation]. University of Ottawa.

Kristjansson, E., Aylesworth, R., Mcdowell, I., & Zumbo, B.D. (2005). Comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement*, *65*(6), 935-953. https://doi.org/10.1177/0013164405275668

Lim, R.G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, *75*(2), 164-174. https://doi.org/10.1037/0021-9010.75.2.164

Lord, F.M. (2012). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates.

Magis, D., Beland, B., & Raiche, G. (2018). difR: Collection of methods to detect dichotomous differential item functioning (DIF). https://cran.r-project.org/web/packages/difR/difR.pdf

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent type I error inflation in differential item functioning. *Educational and Psychological Measurement, 72*(2), 291-311.

Mellenbergh, G.J. (1983). Conditional item bias methods. In S.H. Irvine & J.W. Berry (Ed.), *Human assessment and cultural factors* (pp. 293-302). Springer.

Narayanan, P., & Swaminathan, H. (1994). Performance of the mantel-haenszel and simultaneous item bias procedures for detecting differential. *Applied Psychological Measurement*, *18*(4), 315-328. https://doi.org/10.1177/014662169401800403

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274. https://doi.org/10.1177/014662169602000306

Osterlind, S.J. (1983). *Test item bias*. Sage Publications.

Osterlind, S.J., & Everson, H.T. (2009). *Differential item functioning*. Sage Publications.

Patton, M.Q. (1990). *Qualitative evaluation and research methods.* Sage Publications, Inc.

Price, E.A. (2014). *Item Discrimination, model-data fit, and type I error rates in DIF detection using lord's χ^2, the likelihood ratio test, and the mantel-haenszel procedure* [Doctoral thesis, Ohio University]. OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1395842816

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrica*, *53*(4), 495-502. https://doi.org/10.1007/BF02294403

Rizopoulos, D. (2018). Latent trait models under IRT. https://cran.r-project.org/web/packages/ltm/ltm.pdf

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and mantel-haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116. https://doi.org/10.1177/014662169301700201

Roussos, L.A., & Stout, W.F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and mantel-haenszel type I error performance. *Journal of Educational Measurement*, *33*(2), 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Samuelsen, K.M. (2005). *Examining differential item functioning from a latent class perspective* (3175148) [Doctoral thesis, University of Maryland]. PreQuest.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317-375. https://doi.org/10.3102/10769986006004317

Simon, J.L. (1978). *Basic research methods in social science*. Random House.

Sünbül, Ö., & Ömür Sünbül, S. (2016). Değişen madde fonksiyonunun belirlenmesinde kullanılan yöntemlerde I. tip hata ve güç çalışması [Type I error and power study in methods used to determine differential item functioning]. *Elementary Education Online, 15*(3), 882- 897.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://www.jstor.org/stable/1434855

Vaughn, B.K., & Wang, Q. (2010). DIF trees: Using classification trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6) 941–952. https://doi.org/10.1177/0013164410379326

Wang, W.C., & Su, Y.H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28,* 450–481.

Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement, 37*(4), 316- 335. https://doi.org/10.1177/0146621613476156

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J.R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233–251.

Zwick, R., Thayer, D.T., & Mazzeo, J. (1997). *Describing and categorizing DIF in polytomous items.* Educational Testing Service.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 796–812

https://doi.org/10.21449/ijate.1266808

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Complexity of the Grading System in Turkish Higher Education

Recep Gur[1*], Mustafa Koroglu[2]

[1]Eskişehir Osmangazi University, Faculty of Education, Department of Measurement and Assessment, Eskişehir, Türkiye
[2]Erzincan Binali Yıldırım University, Faculty of Education, Department of Educational Sciences, Erzincan, Türkiye

ARTICLE HISTORY

Received: Mar. 17, 2023
Revised: Nov. 08, 2023
Accepted: Dec. 13, 2023

Keywords:
Higher education,
Grading system,
Fleiss' kappa,
Criterion-referenced assessment,
Norm-referenced assessment.

Abstract: Based on the academic performance grades of university students, various high-stakes decisions are made, including determinations of pass/fail status, the awarding of diplomas, and eligibility for placement in graduate education programs. According to the criteria used, the types of assessment are divided into two assessment, criterion-referenced assessments and norm-referenced assessments. When the grading system of state universities in Turkish higher education is examined, it has been observed that some universities use criterion-referenced assessment, some use norm-referenced assessment, and some use both assessment systems. The purpose of this research is to examine whether inter-university grading systems show significant concordance in the context of university students' letter grades or not. In other words, it is to reveal whether there are skew in the grading systems of public universities. In this context, 250 individuals were simulated in a way that their class/group achievement level would show a normal distribution. Among the public universities in the 2021-2022 Academic Performance Ranking of Universities (URAP), four state universities were determined in the first quarter, second quarter, third quarter, and last quarter. The letter grades of each student's academic success grade in the relevant universities were determined and it was examined whether there was a significant concordance between the letter grades of the students. In the study, it was concluded that in the context of university students' letter grades, inter-university grading systems generally do not show significant concordance. The findings are expected to contribute to the work of the Council of Higher Education and the University Education Commissions.

1. INTRODUCTION

Teaching practices across various levels of the education system aim to equip individuals with skills in cognitive, emotional, and psychomotor domains through activities tailored to specific programs. In the planning of educational and training processes, four fundamental elements of educational programs are considered. These elements include setting objectives, defining content, organizing educational scenarios (learning-teaching processes), and designing measurement and evaluation activities (Demirel, 2007). Measurement and evaluation are crucial for assessing whether the objectives outlined in the educational program align with

*CONTACT: Recep GÜR ✉ verianalizirehberi@gmail.com Eskişehir Osmangazi University, Faculty of Education, Department of Educational Sciences, Ekişehir, Türkiye

individuals' readiness levels, and whether the strategies, methods, and techniques employed in determining their educational status are consistent with their achievements (Atalmış, 2019).

One of the objectives of measurement and evaluation studies in education is to grade students based on their academic achievements in schools (Kubiszyn & Borich, 2000). For this purpose, grades serve as academic performance indicators and play an indispensable role in decisions made during the measurement and evaluation processes. These grades, given to assess students' achievement levels in relation to targeted learning outcomes, can be described as metrics for making judgments about students (Turgut & Baykul, 2015).

These grades can be defined in different ways (4-point system; 100-point system; AA, BA, ..., FF or A1, A2, ..., F3). Higher Education Institutions determine the grading systems they want to implement according to the decisions taken by the university senate. In many countries, while the 100-point system is frequently used in pre-higher education levels, different forms of application are observed in higher education levels. In Turkey's higher education system, lecturers evaluate students' academic achievement through various components. These include test scores, in-class and out-of-class performance, project tasks, and active participation in learning and teaching processes. Lecturers or university senate decides can decide how much weight (weighting percentage) each component will have in grading. For example, 40% of a student's grade is usually midterm exam scores and 60% final scores. It is observed that the regulations regarding the grading of higher education institutions in Turkey are examined, and the grades are generally given in the 4-point system. In addition, some universities include both 100-point and 4-point grading systems in students' transcripts (Özkan, 2016).

Various high-risk decisions (pass/fail, awarding diploma, placement in postgraduate education, etc.) are made based on the results of university students' academic achievement grades. Morever, academic achievement grades are used for admission to a school/programme, promotion to higher grades, graduation and feedback (academic progress for students, quality of education for teachers, status of their children for parents) (Ebel, 1965; Thorndike & Hagen, 1977). Therefore, academic achievement grades can also be effective in shaping the careers and future life of students (Moses & Nanna, 2007). On the other hand, the grades given to the students, regardless of the type of grading system, are the target of some criticisms because they cannot be measured directly. Considering the psychological effects (anxiety, anxiety) that grading systems have on stakeholders all over the world, regardless of letters or numbers, it is quite difficult to say that academic achievement grades are excellent in terms of reliability and validity in determining the level of achievement of individuals (Finkelstein, 1913). The high grades of the students who show the best performance or achieve the aim of the course at the highest level, if they are reliable and valid, can also serve the purpose of increasing students' motivation towards the course (Nitko & Brookhart, 2007). Therefore, grading systems that can evaluate students' performance with minimum error, make precise measurements and provide the opportunity to compare the resulting data should be preferred (Özkan, 2016).

1.1. Criterion-Based Assessments

In universities, assessments conducted to assign letter grades are referred to as "summative assessments," which is one category of assessment based on its intended purpose. On the other hand, the teacher's opinion, the ability of the individual (student), the objectives of the programme, the success level of the group to which the student belongs, the norms developed throughout the country, etc. can be used as criteria in evaluation processes (Turgut, 1983; Airasian, 1994; Haladyna, 1999). One of the critical points of the assessment process is that setting criteria is a very complex and problematic process. The choice of criterion and the amount (level) of the chosen criterion affect the decisions made (Kaya et al., 2017). Determining the criteria may vary according to the teacher's opinion, the success level of the student group who took the exam, the student's ability level, the student's gain development between the end of the program and the beginning of the program, and the program learning outcomes (Martin & Jolly, 2002).

Criterion-based assessments fall into two categories. These are criterion referenced (absolute) assessment and norm-referenced assessment. In criterion-referenced assessment, before the measurement process is performed, the proficiency standard is determined and the success of the student is evaluated independently of the group's performance; in norm-referenced assessment, after the measurement process is made, the success of the student is evaluated according to the relative criterion obtained based on the success grades of the class/group (Kubiszyn & Borich, 2000; Atılgan et al., 2011). It is aimed to evaluate student performances norm-referenced to each other by determining norm-referenced criteria in line with the arithmetic mean and standard deviation scores based on the performances of the students participating in the exam. As a result of the assessment, students can get an idea about their learning levels and meeting the expectations of the students in this process can motivate them (Kaysi et al., 2017).

In the norm-referenced assessment, since the success of the student is determined by the success position among the other students in the class, a student with a good position in a group with a low average success rate may receive a high letter grade or not fail the course. As for criterion-reference assessment, regardless of whether the group's performance is low or high, if the student does not meet the proficiency standard or standard set determined before the measurement process, the student is considered unsuccessful (Kubiszyn & Borich, 2000; Özçelik, 1992). In this system, considering the standards set by the lecturers or institutions, students who reach the relevant standards are considered successful (Mandernach, 2003). Academic achievement grades determined within the scope of criterion-referenced assessment may depend on some factors arising from the student and the lecturer. Among the factors affecting students' academic achievement grades are the faculty members' ability to convey information, their ability to create a suitable psychological environment for students, their psychological state during the evaluation process, etc. Some student-dependent variables, likewise, the achievement level of the students, the approaches towards the lecture, and the lecturer, can also significantly affect the academic achievement grades of the students.

1.2. Differences in the Evaluation System Between Universities

Regulations regarding the assessment systems/grading system used by universities may differ. Beyond the differences in the assessment system between universities, there may be differences even in different faculties of the relevant university (Atılgan et al., 2012). Some universities may use complex scoring systems by including lettering and percentage systems etc. together in their regulations. Although the usage methods of grading systems differ on the basis of countries, the Higher Education Law No. 2547 is taken into account when determining the education, examination and grading systems in universities in Turkey. Within the scope of the law, it is stated that "The education and training carried out according to the characteristics and needs of the establishment in higher education institutions and the principles related to the diplomas which are awarded based on this are specified in the education and examination regulations to be prepared by each university". Higher education institutions based on this law make judgments about student achievements by using different grading systems and passing grades (Özkan, 2016). In many universities, norm-referenced or criterion-referenced grading systems are implemented using a grade range based on T score (Atalmış, 2019). Öztürk-Gübeş (2021) states that in grading systems, when calculating the composite score by combining different evaluations, multiplying the midterm and final raw scores with a ratio is a separate problem. Although the final and midterm assessments have different standard deviations and means, they are assumed to be on the same scale. The target weight of a score component and its real impact on grades can be very different. The stated reasons prevent the grades of individuals who graduated from different higher education institutions from being comparable, but they may also cause measurement bias for students. Bias is defined as a systematic error that leads to the advantage or disadvantage of one group (Reynolds, Livingston, & Wilson, 2006). For example, considering two different university grading systems where it is easy and difficult to get an AA grade, when the grades are not comparable, this situation causes injustice and measurement bias problems arising from grading systems. The lack of a standard system in higher education institutions shows that the grades of university students who continue their education in different higher education institutions, which provide indicators of their academic success, can find direction in different values. The fact that universities and the Higher Education Council have different grade conversion tables-based on different passing grades and evaluation practices-complicates the matching of grades both within and between institutions. This lack of standardization creates issues in comparing the academic success levels of students, even if they graduate from the same programs.

Although the objectives such as determining the level of access of individuals to the attainments required to be achieved in the curriculum throughout the world, removing the individuals who fall below the specified standards from the system, and increasing the quality of education, make criterion-referenced assessment more common, higher education institutions in Turkey prefer to use norm-referenced assessment more. There are studies in the literature on how to use criterion-referenced or norm-referenced systems. Many studies in the literature focused on comparing criterion-referenced and norm-referenced rating systems (Başol- Mandernach, 2003; Göçmen, 2004; Nartgün, 2007; Duman, 2011; Lok et al., 2016; Özkan, 2016; Sayın, 2016; Atalmış, 2019). In addition to these, there are several studies examining the errors made in the weighting of the activities that are the subject of passing grades in the norm-referrenced assessment (Kelley & Zarembka, 1968; Tinkelman et al., 2013; Öztürk-Gübes, 2021). Although there are studies in the literature in which different grade-taking systems are compared and the advantages and disadvantages of criterion-referenced and norm-referrenced assessment systems are stated, the class averages and standard deviations are controlled with simulative data equally, and the letter of each student's academic achievement raw score in different quarters according to the URAP ranking of the universities. No study has been found on the concordance between the grading systems in higher education institutions/letter grades of students. The aim of this research is to investigate the level of concordance between different universities' grading systems, particularly concerning students' letter grades. In other words, it is to reveal whether there are skews in the grading systems of public universities.

2. METHOD

2.1. Research Model

This research aims to determine the letter grade equivalents of each student's academic achievement grade generated by the Monte Carlo method in the relevant universities and to examine the coefficients of concordance between the letter grades of the students. Therefore, this research is a Monte Carlo simulation study that seeks to answer the question "What would happen if each student's academic achievement grade was like this?" (Dooley, 2002).

2.2. Generating Data

Ethics committee decision is not required in this study, since the analysis was conducted on simulative data. In order to apply a norm-referrenced assessment, the grade distribution of the group should show a normal distribution. Due to the difficulty of achieving a normal distribution of academic grades in a small sample size, this study used simulated data for 250 individuals. The data were modeled to have a normal distribution in class/group achievement level with a mean (*M*) of 62.55 and a standard deviation (*SD*) of *12.53.*

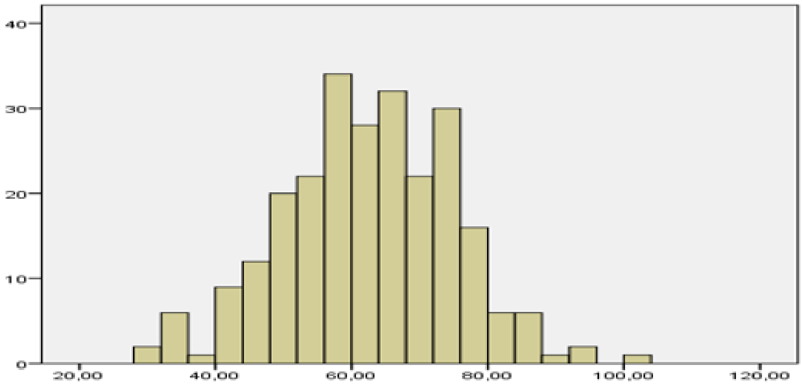
When the class achievement levels are examined according to the limit values of the raw achievement grade point averages (GPA) in the grading systems of the universities, the arithmetic average is around 62.50, since the class raw achievement GPA range of 60-65 is generally defined as the "middle" class level. Thus, in the Monte Carlo simulation process, care was taken to ensure that the simulated data represented real-world scenarios by using parameters to simulate 250 individuals with *M=62.55, SD=12.53*, whose class/group achievement level was normally distributed. In Monte-carlo simulation studies, a data set is created in accordance with the conditions specified by the researcher. Monte-carlo simulation studies diversify the data sets and provide an effective and fast comparison between grading systems etc. Simulative data were produced by using SimulCAT (Han, 2011) software. The descriptive statistics regarding the academic achievement scores of the students are given below.

Table 1. *Descriptive statistics on students' academic achievement scores.*

N	M	Median	Mode	SD	Min	Max	Skewness	SE_{skewness}	Kurtosis	SE_{Kurtosis}
250	62.55	62.26	56.35	12.53	28.71	100	-.08	.15	.03	.31

When Table 1 is examined, there are 250 students and their academic success grades are distributed between *28.71* and *100* (*M=62.55, SD=12.53*). In addition, it can be stated that the arithmetic mean, median and mode values are close to each other, the kurtosis and skewness coefficients are close to zero, the ratio of the skewness and kurtosis coefficients to the standard error is close to zero, the data show a normal distribution curve and the basic basis of norm-referrenced assessment is provided. The distribution of students' academic achievement scores is given in Figure 1.

Figure 1. *Distribution of students' academic achievement scores*



When Figure 1 is examined, it shows the normal distribution curve characteristic of students' academic achievement scores.

2.3. Application Process

There are 120 state and 59 foundation universities among 179 universities in the Academic Performance Ranking of Universities in Turkey 2021-2022 (URAP). Among the URAP 2021-2022 public universities, Istanbul University (5th), Harran University (58th), Bartın University (90th), Kırklareli University (152nd) and four state universities were determined in the first quarter, second quarter, third quarter and last quarter, respectively (URAP, 2022).

In the grading system at Istanbul University, students with a raw achievement score (RAS) below 35 are automatically given an FF. On the other hand, students with an RAS of 100 are excluded from the Norm-Referenced Assessment System and receive an AA directly.Therefore, after examining the simulated data set between *28.71* and *100* (*M=62.55, SD=12.53, N=250*) regarding student academic achievement grades, the letter grades of the students whose RAS were below *35* were directly FF; The arithmetic mean and standard deviation were calculated and found (*M=63.38, SD=11.27*) in order to calculate the T scores of the other students, without directly defining the letter grades of the students with an RAS of 100 and not including the raw grades of the related students in the Norm-Referenced Assessment System. Table 2 shows that class achievement level at Istanbul University (*M=63.38*) was defined as "Above Average" ($63\leq\mu<71$) according to the limit values of raw grade point averages. Since there are N≥20, *SD* (σ)≥8, the assessment system at Istanbul University uses variable intervals based on the mean and standard deviation below (Istanbul University, 2022).

Table 2. *Variable intervals method in Istanbul University norm referenced assessment system.*

Letter Grade	Very Poor/Very Low: μ<44	Poor/Low: 44≤μ<50	Below Average: 50≤μ<56
AA	$[\mu+1,881\sigma, 100]$	$[\mu+1,645\sigma, 100]$	$[\mu+1,476\sigma, 100]$
BA	$[\mu+1,405\sigma, \mu+1,881\sigma)$	$[\mu+1,175\sigma, \mu+1,645\sigma)$	$[\mu+0,994\sigma, \mu+1,476\sigma)$
BB	$[\mu+0,706\sigma, \mu+1,405\sigma)$	$[\mu+0,524\sigma, \mu+1,175\sigma)$	$[\mu+0,358\sigma, \mu+0,994\sigma)$
CB	$[\mu+0,332\sigma, \mu+0,706\sigma)$	$[\mu+0,126\sigma, \mu+0,524\sigma)$	$[\mu-0,075\sigma, \mu+0,358\sigma)$
CC	$[\mu-0,176\sigma, \mu+0,332\sigma)$	$[\mu-0,468\sigma, \mu+0,126\sigma)$	$[\mu-0,772\sigma, \mu-0,075\sigma)$
DC	$[\mu-0,643\sigma, \mu-0,176\sigma)$	$[\mu-0,878\sigma, \mu-0,468\sigma)$	$[\mu-1,126\sigma, \mu-0,772\sigma)$
DD	$[\mu-1,175\sigma, \mu-0,643\sigma)$	$[\mu-1,405\sigma, \mu-0,878\sigma)$	$[\mu-1,645\sigma, \mu-1,126\sigma)$
FF	$[35, \mu-1,175\sigma)$	$[35, \mu-1,405\sigma)$	$[35, \mu-1,645\sigma)$
Letter Grade	Average: 56≤μ<63	Above Average: 63≤μ<71	Good/High: 71≤μ<80
AA	$[\mu+1,227\sigma, 100]$	$[\mu+0,915\sigma, 100]$	$[\mu+0,583\sigma, 100]$
BA	$[\mu + 0,739\sigma, \mu + 1,227\sigma)$	$[\mu + 0,385\sigma, \mu + 0,915\sigma)$	$[\mu + 0,100\sigma, \mu + 0,583\sigma)$
BB	$[\mu+0,126\sigma, \mu+0,739\sigma)$	$[\mu-0,075\sigma, \mu+0,385\sigma)$	$[\mu-0,305\sigma, \mu+0,100\sigma)$
CB	$[\mu- 0,358\sigma, \mu + 0,126\sigma)$	$[\mu-0,524\sigma, \mu-0,075\sigma)$	$[\mu-0,739\sigma, \mu-0,305\sigma)$
CC	$[\mu-0,878\sigma, \mu-0,358\sigma)$	$[\mu-0,994\sigma, \mu-0,524\sigma)$	$[\mu-1,126\sigma, \mu-0,739\sigma)$
DC	$[\mu-1,227\sigma, \mu-0,878\sigma)$	$[\mu-1,341\sigma, \mu-0,994\sigma)$	$[\mu-1,476\sigma, \mu-1,126\sigma)$
DD	$[\mu-1,751\sigma, \mu-1,227\sigma)$	$[\mu-1,881\sigma, \mu-1,341\sigma)$	$[\mu-2,054\sigma, \mu-1,476\sigma)$
FF	$[35, \mu-1,751\sigma)$	$[35, \mu-1,881\sigma)$	$[35, \mu-2,054\sigma)$
Letter Grade	Very Good/Very High: μ≥80		
AA	$[\mu+0,440\sigma, 100]$		
BA	$[\mu- 0,100\sigma, \mu + 0,440\sigma)$		
BB	$[\mu-0,496\sigma, \mu-0,100\sigma)$		
CB	$[\mu-0,915\sigma, \mu-0,496\sigma)$		
CC	$[\mu-1,282\sigma, \mu-0,915\sigma)$		
DC	$[\mu-1,645\sigma, \mu-1,282\sigma)$		
DD	$[\mu-2,326\sigma, \mu-1,645\sigma)$		
FF	$[35, \mu-2,326\sigma)$		

When the grading system of Harran University is examined, students whose raw achievement score is below 35 are given FF directly; Students with an RAS of 90 and above are not included in the Norm-Referenced Assessment System by taking AA directly. When the grading system of Harran University (Harran University, 2022) is examined, the raw achievement score (RAS) is directly FF; Students with an RAS of 90 and above are not included in the Norm-Referenced Assessment System by taking AA directly. Therefore, the simulated data set between *28.71* and *100* (*M=62.55, SD=12.53, N=250*) related to student academic achievement scores was examined and the letter grades of the students whose RAS was below *35* were directly FF; The letter grades of the students with an RAS of *90* and above were directly defined as AA, and the arithmetic mean and standard deviation were calculated (*M=63.11, SD=10.76*) in order to calculate the T scores of the other students without including the raw achievement scores of the related students in the Norm-Referenced Assessment System. Table 3 indicates that, class achievement level at Harran University (*M=63.11*) was defined as "*Good*" (60≤*μ*<70) according to the limit values of raw grade point averages. Since there are N≥20, *SD*≥8, the assessment system at Harran University uses variable intervals based on the mean and standard deviation below (Harran University, 2022):

Table 3. *Variable intervals method in Harran University norm referenced assessment system.*

Letter Grade	Poor (μ < 50)	Average (50 ≤ μ < 60)	Good (60 ≤ μ <70)
AA	$[\mu + 3.00, 100]$	$[\mu + 2.30\sigma, 100]$	$[\mu + 1.50\sigma, 100]$
BA	$[\mu + 2.60\sigma, \mu + 3.00\sigma)$	$[\mu + 1.90\sigma, \mu + 2.30\sigma)$	$[\mu + 1.10\sigma, \mu + 1.50\sigma)$
BB	$[\mu + 2.20\sigma, \mu + 2.60\sigma)$	$[\mu + 1.50\sigma, \mu + 1.90\sigma)$	$[\mu + 0.70\sigma, \mu + 1.10\sigma)$
CB	$[\mu + 1.30\sigma, \mu + 2.20\sigma)$	$[\mu + 0.80\sigma, \mu + 1.50\sigma)$	$[\mu + 0.30\sigma, \mu + 0.70\sigma)$
CC	$[\mu +0.40\sigma, \mu + 1.30\sigma)$	$[\mu + 0.10\sigma, \mu + 0.80\sigma)$	$[\mu - 0.10\sigma, \mu + 0.30\sigma)$
DC	$[\mu - 0.30\sigma, \mu +0.40\sigma)$	$[\mu - 0.80\sigma, \mu +0.10\sigma)$	$[\mu - 1.30\sigma, \mu - 0.10\sigma)$
DD	$[\mu - 1.00\sigma, \mu - 0.30\sigma)$	$[\mu - 1.70\sigma, \mu - 0.80\sigma)$	$[\mu - 2.50\sigma, \mu - 1.30\sigma)$
FF	$<\mu - 1.00\sigma$	$<\mu - 1.70\sigma$	$<\mu - 2.50\sigma$
Letter Grade	Very Good (70 ≤ μ <80)	Excellent (μ ≥80)	
AA	$[\mu + 1.00\sigma, 100]$	$[\mu + 0.50\sigma, 100]$	
BA	$[\mu + 0.65\sigma, \mu + 1.00\sigma)$	$[\mu + 0.20\sigma, \mu + 0.50\sigma)$	
BB	$[\mu + 0.30\sigma, \mu + 0.65\sigma)$	$[\mu - 0.10\sigma, \mu + 0.2\sigma)$	
CB	$[\mu - 0.05\sigma, \mu + 0.30\sigma)$	$[\mu - 0.40\sigma, \mu -0.10\sigma)$	
CC	$[\mu - 0.40\sigma, \mu - 0.05\sigma)$	$[\mu - 0.70\sigma, \mu - 0.40\sigma)$	
DC	$[\mu - 1.70\sigma, \mu - 0.40\sigma)$	$[\mu - 2.10\sigma, \mu - 0.70\sigma)$	
DD	$[\mu - 3.00\sigma, \mu - 1.70\sigma)$	$[\mu - 3.50\sigma, \mu - 2.10\sigma)$	
FF	$<\mu - 3.00\sigma$	$<\mu - 3.50\sigma$	

250 individuals were simulated with a normal distribution of grade achievement level (M=*62.55, SD=12.53; Min=28.71, Max=100*). In line with these data, when the grading system of Bartın University (Bartın University, 2022) is examined, students with a RAS below 15 are not included in the Norm-Referenced Assessment System. Since the minimum RAS in the simulated study group was 28.71, the RAS of all students was included in the Norm-Referenced Assessment System. Therefore, the location parameters (*M=62.55, SD=12.53*) did not change in order to calculate the T scores of the students. In addition, as stated in the grading system of Bartın University, the RAS is directly defined as FF for students whose RASis below threshold limit 45. Table 4 shows that class achievement level at Bartın University (*M=62.55*) was defined as '*Average*' (50≤*μ*<65) according to the limit values of raw grade point averages. It is assumed that students' end-of-term (final) exam raw academic scores are at least 50. "Limit

scores of the criteria applied in determining the class achievement level" and "RAS limit values of letter grades according to class achievement level at Bartın University" are given below (Bartın University, 2022):

Table 4. *Limit scores of the criteria applied in determining the class achievement level.*

Criterion	Class Achievement Level				
	Very Poor	Poor	Average	Good	Very Good
Class RAS average lower limit	0	35	50	65	85
Class RAS average upper limit	34.99	49.99	64.99	84.99	100
Limit of inclusion in assessment	15	15	15	20	20
Threshold limit of RAS	45	45	45	50	50
End-of-term (final) exam RAS limit	50	50	50	60	60

Table 5. *RAS limit values of letter grades according to class achievement level at Bartın University.*

Letter Grades	Class Achievement Level				
	Very Poor	Poor	Average	Good	Very Good
AA	75	80	85	90	95
BA	70	70	75	80	85
BB	65	65	65	70	80
CB	60	60	60	65	75
CC	55	55	55	60	70
DC	50	50	50	55	65
DD	45	45	45	50	50
FF	<45	<45	<45	<50	<50

When the grading system of Kırklareli University (Kırklareli University, 2022) is examined, the norm-referenced assessment limit is 20 and students who fall below 20 are not included in the norm-referenced assessment system. 250 individuals were simulated with a normal distribution of grade achievement level (M=*62.55*, *SD=12.53; Min=28.71*, *Max=100*). In line with these data, since the minimum RAS in the simulated study group was *28.71*, the RAS of all students was included in the norm-referenced assessment system. Therefore, the location parameters (*M=62.55*, *SD=12.53*) did not change in order to calculate the T scores of the students. In addition, as stated in the grading system of Kırklareli University, students who score below the lower limit of success grade 40 points will be considered unsuccessful, and students who fall below 40 are directly defined as FF. Table 6 indicates that at class achievement level Kırklareli University (*M=62.55*) is defined as *'Very good'* (62.5≤μ<70) according to the limit values of raw grade point averages. Norm referenced assessment system Kırklareli Unıversity is given below (Kırklareli University, 2022):

Table 6. *Norm referenced assessment system of Kırklareli Unıversity.*

Class Level	Intervals over 100 (Class average)		Lower limits of norm refernced grades according to T-scores								
	Lower Limit	Upper Limit	AA (4.00)	BA (3.50)	BB (3.00)	CB (2.50)	CC (2.00)	DC (1.50)	DD (1.00)	FD (0.50)	FF (0.00)
Overachievement	80.00	100.00	57	52	47	42	37	32	27	22	<22
Excellent	70.00	79.99	59	54	49	44	39	34	29	24	<24
Very Good	62.50	69.99	61	56	51	46	41	36	31	26	<26
Good	57.50	62.49	63	58	53	48	43	38	33	28	<28
Above Average	52.50	57.49	65	60	55	50	45	40	35	30	<30
Average	47.50	52.49	67	62	57	52	47	42	37	32	<32
Poor	42.50	47.49	69	64	59	54	49	44	39	34	<34
Bad	0	42.49	71	66	61	56	51	46	41	36	<36

Although the class averages and standard deviations are simulated, the study finds variations in the grading systems of different universities. Specifically, the terms “Above Average”, “Good”, “Average”, and “Very Good” are defined differently across the four state universities examined.

2.4. Data Analysis

The letter grades of each student's academic achievement score/ RAS at the relevant universities were determined and whether there was a significant concordance between the letter grades of the students was examined by Cohen's kappa coefficient and Fleiss' kappa coefficient. While calculating the coefficient of agreement between the two evaluators/universities, Cohen's kappa coefficient is used; Fleiss' kappa coefficient is used in cases where the agreement between more than two raters is measured (Fleiss, 1971). The STATA 14 program was used to calculate Cohen's kappa coefficient between two universities and Fleiss' kappa coefficient between four universities. Kappa coefficients are suggested to be interpreted as follows (Landis & Koch, 1977; Fleiss, 1981):

Table 7. *The value ranges for Kappa coefficients.*

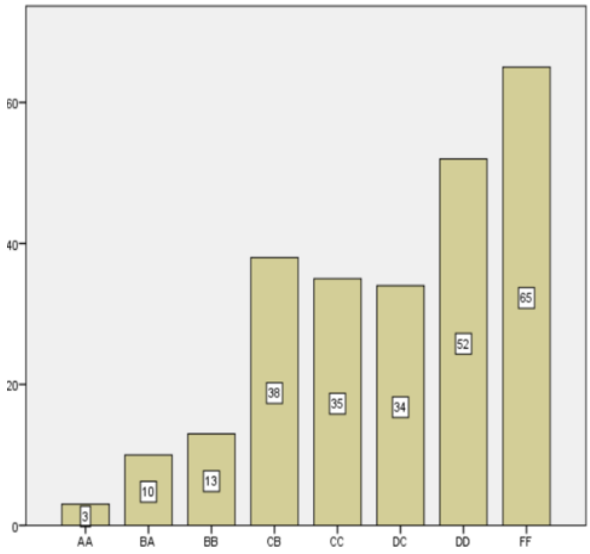
Coefficient	Interpretation
<.00	Poor
.00 to .20	Slight
.21 to .40	Fair
.41 to .60	Moderate
.61 to .80	Substantial
.81 to 1.00	Almost Perfect

Kappa coefficients are interpreted as “*poor*”, “*slight*”, “*fair*”, “*moderate*”, “*substantial*” and “*almost perfect*” respectively.

3. FINDINGS

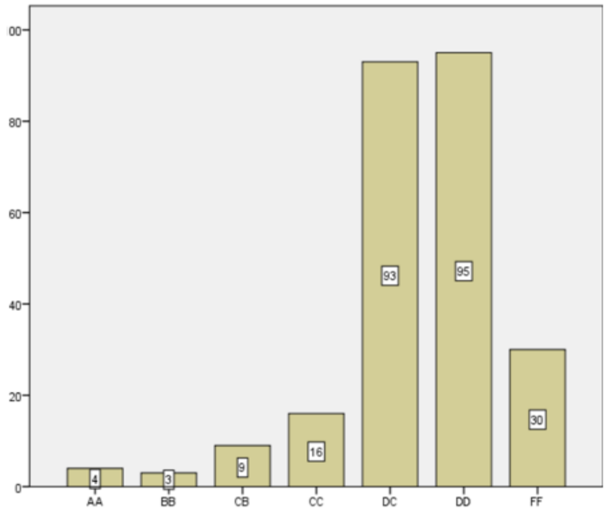
The distribution of the raw achievement score of each simulated student according to the grading system of Istanbul University regarding the letter grades of the students is presented in Figure 2.

Figure 2. *Distribution of students' letter grades according to the Istanbul University grading system.*



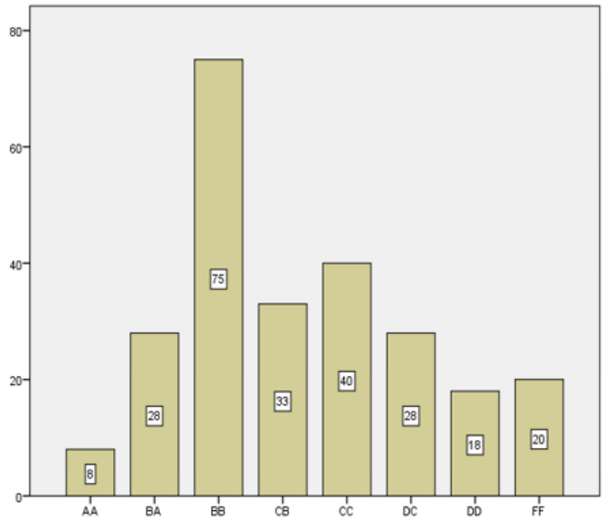
According to the grading system of Istanbul University, the number of students who received high letter grades AA (*f*=3), BA (*f*=10), and BB (*f*=13) is low. Most students have received DD (*f*=52) and FF (*f*=65) grades. In other words, student letter grades generally piled up to unsuccessful/low letter grades. The distribution of each student's raw achievement score according to the grading system of Harran University regarding the letter grades of the students is presented in Figure 3.

Figure 3. *Distribution of students' letter grades according to the Harran University grading system.*



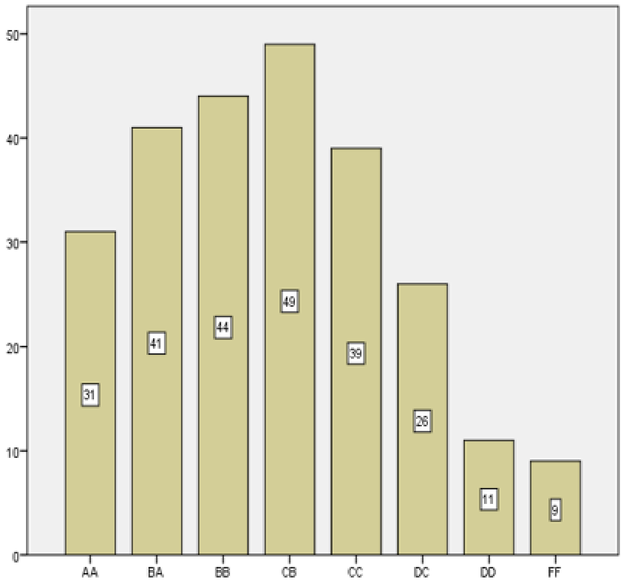
When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of Harran University, it is seen that the number of students who receive AA (*f=4*), BB (*f=3*), CB (*f=9*) and CC (*f=16*) is low. It has been observed that students generally piled up on letter grades of DD (*f=93*), DD (*f=95*) and FF (*f=30*). In other words, student letter grades and letter grades have generally piled up on unsuccessful letter grades. Additionally, despite a class of 250 students, when the letter grade levels of the students are examined, it is notable that the grades of the students de-escalated (scree) from AA level to BB level, and that there was no student who received a "BA" letter grade. The distribution of the student's raw achievement score according to the grading system of Bartın University regarding the letter grades of the students is given in Figure 4.

Figure 4. *The distribution of hidden letter grades according to Bartın University grading system.*



When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of Bartın University, it can be stated that while the number of students with a letter grade of BB and above is high, the number of students with a letter grade of DC or below is low. In other words, student letter grades are generally piled up successful/high letter grades. The distribution of each student's raw achievement score according to the grading system of Kırklareli University regarding the letter grades of the students is given in Figure 5.

Figure 5. *Distribution of students' letter grades according to the Kırklareli University grading system.*



When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of Kırklareli University, it can be stated that while the number of students with a letter grade of CB and above is high, the number of students with a letter grade of DC or below is low. In other words, student letter grades generally piled up on successful/high letter grades. Distribution of students' letter grades according to the grading system of four different state universities (Istanbul University Q1, Harran UniversityQ2, Bartın University Q3, Kırklareli University Q4) ranked in the first quarter, second quarter, third quarter and last quarter, respectively, among URAP 2021-2022 public universities given in Table 8.

Table 8. *Distribution of students' letter grades according to the grading system of universities.*

	Istanbul University	Harran University	Bartın University	Kırklareli University
AA	3	4	8	31
BA	10	3	28	41
BB	13	-	75	44
CB	38	9	33	49
CC	35	16	40	39
DC	34	93	28	26
DD	52	95	18	11
FF	65	30	20	9

When the letter grade equivalents of each student's raw achievement score are examined according to the grading system of the universities, the number of students who received AA letter grades increased as we move across Istanbul University to Kırklareli University (from Q1 to Q4); The number of students who received FF letter grades decreased. For example, while there are *3* students with AA and *65* students with FF at Istanbul University; Kırklareli University has *31* students with AA and *9* students with FF. In other words, student letter grades at Bartın and Kırklareli University have generally piled up on successful/high letter grades. When the letter grade coefficients of the universities are examined, they equal to *AA-4.00, BA-3.50, BB-3.00, CB-2.50, CC-2.00, DC-1.50, DD-1.00, FF-0.00* respectively. The difference in letter grade coefficients between universities is given in Table 9.

Table 9. *The difference in letter grade coefficients between universities.*

Letter Grade Coefficient Differences	Istanbul-Harran	Istanbul-Bartın	Istanbul-Kırklareli	Harran-Bartın	Harran-Kırklareli	Bartın-Kırklareli
-2.50	-	-	-	-	-	-
-2.00	-	-	19 %7.60	4 %1.60	55 %22	-
-1.50	-	50 %20	88 %35.20	111 %44.40	104 %41.60	-
-1.00	35 %14	125 %50	121 %48.40	75 %30	62 %24.80	11 %4.40
-0.50	1 %0.4	47 %18.80	10 %4.00	28 %11.20	16 %6.40	123 %49.20
0.00	111 %44.40	28 %11.20	12 %4.80	32 %12.80	13 %5.20	116 %46.40
0.50	55 %22	-	-	-	-	-
1.00	48 %19.20	-	-	-	-	-

Although the class averages and standard deviations are equal, it has been observed that there are differences of up to 2 coefficients between the letter grades of the students in different universities with the same raw score. For example, when the letter grade equivalents of students between raw achievement scores from *70.137* to *71.535* are examined, it corresponds to the letter grades of *CC* in Istanbul University, *DC* in Harran University, *BB* in Bartın University and *BA* in Kırklareli University. It is notable that the letter grades corresponding to the raw achievement score are different from each other, although the raw achievement score is between *70.137* and *71.535*, the class mean and standard deviations of all four state universities are controlled. The kappa coefficients showing the concordance between the grading systems of the universities are presented in Table 10.

Table 10. *Kappa coefficients between the grading systems of universities.*

Universities	Kappa	*p*
İstanbul-Harran (Q_1 x Q_2)	.325^a	.00*
İstanbul-Bartın (Q_1 x Q_3)	-.002^a	.92
İstanbul-Kırklareli (Q_1 x Q_4)	-.059^a	.00*
Harran-Bartın (Q_2 x Q_3)	.034^a	.07
Harran-Kırklareli (Q_2 x Q_4)	-.031^a	.06
Bartın-Kırklareli (Q_3 x Q_4)	.374^a	.00*
Q_1 x Q_2 x Q_3 x Q_4	.08^b	.00*

a : Cohen Kappa coefficient
b : Fleiss' Kappa coefficient
* : p<.05

When Table 5 is examined, it is seen that in terms of letter grades of university students, inter-university grading systems generally do not show significant concordance and Kappa coefficients are poor (below .00), slight (between .00 and .20) or fair (between .20 and .40) level was found.

4. DISCUSSION and CONCLUSION

In this Monte Carlo simulation study, it was analysed whether there is a significant concordance between the letter grade equivalents of students' raw academic achievement scores. The study focuses on four public universities Istanbul University, Harran University, Bartın University, and Kırklareli University ranked in different quarters of the 2021-2022 URAP Academic Performance Ranking. As a result of the research, as one moves from the first quarter to the last quarter among the 2021-2022 URAP state universities, the number of students with *AA* letter grades increased; it was concluded that the number of students who received *FF* letter grades decreased.

Although the class averages and standard deviations of the students in different universities are controlled by simulative data according to the limit values of the raw grade point averages in the grading systems of the universities, they are classified as "*Above Average"*, *"Good"*, *"Average"* and "*Very Good"*, respectively. All four of the four state universities have different definitions of the respective class achievement level. In addition, although the class averages and standard deviations are equal, the observation of differences up to *2* coefficients between the letter grades of students in different universities with the same raw score reveals the skewness between the grading systems of state universities. In the context of university students' letter grades, it has been concluded that inter-university grading systems generally do not show significant concordance and kappa coefficients are poor, slight or fair level.

Research findings show that there is generally no concordance between the grading systems of public universities. Although it is seen that this situation arises from the difference in the norm-referenced assessment algorithms used by the universities discussed in the study, it can be thought that it creates a bias in favor of the students studying at some universities that use thenorm-referencedassessment system. The difference between the criterion-referenced and norm-referenced assessment systems, as well as the injustice caused by the differences between the norm-referenced assessment systems in practice, directly affect the applications for graduate education or lateral transfer applications of graduate university students with their undergraduate graduation averages.

According to the general results of the study, the finding of differences between student letter grades stemming from the grading systems in the universities discussed in the research in the process of determining letter grades is similar to the studies in which students' letter grades are compared using different systems (criterion-referencedor norm-referenced assessment) (Mandernach, 2003; Başol-Göçmen, 2004; Nartgün, 2007; Duman, 2011; Lok *et al.*, 2016; Özkan, 2016; Sayın, 2016; Atalmış, 2019). While Lok *et al.* (2016) emphasize that criterion-referenced and norm-referencedassessment systems should be compatible and complementary; Kaya and Semerci (2017) receive opinions from lecturers about the positive and negative aspects of criterion-referenced and norm-referenced assessment systems. Sayın (2016) and Atalmış (2019) reached the conclusion that norm-referencedassessment received higher letter grades than criterion-referenced assessment, and that measures should be taken against the negativities of using the norm-referenced assessment system in Başol Göçmen (2004) and Mandernach (2003) studies. Duman (2011) stated as a result of the research that prospective classroom teachers have negative perceptions towards norm-referenced assessment. Differing from these findings, Atılgan *et al.* (2012) and Nartgün (2007) have stated that grading by using the criterion-referenced assessment system is a more accurate practice, however, Atılgan *et al.*

(2012) concluded that in a study where the norm-referenced assessment and criterion-referenced assessment are used to compare the obtained letter grades, in the case of a norm-referencedassessment, student letter grades will be approximately *40%* less than the letter grades obtained as a result of the criterion-referenced assessment.

When the studies showing that norm-referencedassessment increases student grades/causes grade inflation are examined, it can be explained that using norm-referenced assessment is advantageous in terms of instructors' inability to prepare questions in accordance with the principles of assessment and assessment, reducing the errors caused by the assessment tool, and not punishing the student for failure that may arise from lack of teaching. Turgut and Baykul (2015) state that in cases where the group distribution in the norm-referenced assessment is normal, the letter grades to be obtained will also be symmetrical, and in other cases, the letter grades to be taken will be more affected by the extreme values. Thorndike (2005) emphasizes that while preparing tests with appropriate psychometric properties, the difficulty levels of the items should be balanced and *25%* of the items should be above medium difficulty, *50%* moderate and *25%* below medium difficulty.

There have also been several studies (Kelley & Zarembka, 1968; Öztürk-Gubes, 2021; Tinkelman *et al*., 2013) examining errors in weighting of activities that are subject to passing grades in norm-referenced assessment. Öztürk-Gübes (2021) emphasized that the agreement between the grade values obtained by weighting according to the raw scores and the grade values that were weighted after standardization changed, and the fit between the grade values calculated by both methods decreased as the difference between the standard deviations of the midterm and final measurements increased. As for Özkan (2016), in the study titled 'Chaos in university graduation grades and conversion tables' concluded that The Council of Higher Education's grade conversion table provides a transformation in favor of the students in the universities with 50 passing grades, and against the students in the universities where 60 and 70 passing grades are applied. Özkan (2016) stated that the determination of the passing grades and the systems of the students who graduated from different higher education institutions cause problems in the grade conversion. He stated that the determination of the passing grades and the systems of the students who graduated from different higher education institutions cause problems in the grade conversion.

In line with these discussions, it shows that the grading systems of higher education institutions, which are applied in different ways based on the education and examination regulations specified in the higher education law numbered 2547 and prepared by the relevant commissions of different universities, and the determination of passing grades cause various problems. The lack fit between the grading systems of universities prevents the grades of individuals who graduated from different higher education institutions from being comparable, but also undermines the validity of the measurement results as it will cause measurement bias in favor of students in some universities. Students with an equivalent bachelor's degree are expected to practice an equivalent profession. Nevertheless, it is demonstrated as a crucial problem that the skewness caused by the grading systems of the universities from which the students graduated need to be considered and solved.

According to the academic achievement grades of university students, various high-risk decisions such as pass/fail, awarding diplomas, placement in graduate education, acceptance for transfer, etc. are made. Considering that these grades are effective in shaping students' careers and future lives, it is expected that there should be a standard grading system to ensure fairness among Higher Education Institutions. In order to eliminate the skewness between the grading systems of universities, the Education Commissions, which regulate the grading systems of universities, under the leadership of the Council of Higher Education, organize workshops, panels, etc., it is recommended to organize programs and make the necessary arrangements for the standard grading system to serve its purpose, taking into account the qualifications expected from the graduates of the relevant faculty. The findings are expected to contribute to the work of the Council of Higher Education and the education commissions that regulate the grading systems of universities.

Acknowledgments

The authors would like to thank the blind reviewers for their valuable suggestions. While this study was under development, an abstract of some parts of the study was presented as an oral (abstract) paper titled "*The complexity of grading system in Turkish higher education*" by Gür, R. and Köroğlu M. at the VIIIth International TURKCESS Education and Social Sciences Congress held in Hatay on 21 July 2022.

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**: Erzincan Binali Yıldırım University, Human Research and Educational Sciences Ethics Committee, 30/12/2022-12/06.

Authorship Contribution Statement

Recep Gur: Investigation, Resources, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Mustafa Koroglu:** Literature Review, Methodology, Supervision, and Writing-original draft.

Orcid

Recep Gur https://orcid.org/0000-0002-3686-4199
Mustafa Koroglu https://orcid.org/0000-0001-9610-8523

REFERENCES

Airasian, P.W. (1994). *Classroom assessment*. Mc Graw Hill. Inc. New York.

Atalmış, E.H. (2019). A statistical comparison of norm-referenced assessment systems usepar in higher education in Turkey. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(1), 12-29. https://doi.org/10.21031/epod.487335

Atılgan, H., Yurdakul, B., & Öğretmen, T. (2012). A research on the relative and absolute evaluation for determination of students achievement. *Inonu University Journal of the Faculty of Education*, *13*(2), 79-98.

Atılgan, H., Kan, A., & Doğan, N. (2011). *Eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in education*]. Anı Yayıncılık.

Bartin University (2022). Bartın Üniversitesi Bağıl Değerlendirme Sistemi Uygulama Yönergesi [Bartin University Relative Assessment System Implementation Instruction]. https://kms.kaysis.gov.tr/Home/Kurum/85269548

Basol Gocmen, G. (2004). *Değerlendirmeye genel bir bakış: Kriter-referanslı (mutlak) ya da norm-referanslı (bağıl) değerlendirme* [*An overview of evaluation: Criterion-referenced (absolute) or norm-referenced (relative) evaluation*]. XIII. Ulusal Eğitim Bilimleri Kurultayı'nda sunulmuş bildiri. İnönü Üniversitesi Eğitim Fakültesi, Malatya.

Demirel, Ö. (2007). *Eğitimde program geliştirme* [*Program development in education*]. Pegem A Yayıncılık.

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). Blackwell.

Duman, B. (2011). Sınıf öğretmeni adaylarının bağıl değerlendirmeye ilişkin görüşleri [Opinions of classroom teacher candidates regarding relative assessment]. *NWSA-E Journal of New World Sciences Academy*, *6*(1), 536-548

Ebel, R.L. (1965). *Measuring educational achievement*. Prentice-Hall, Inc.

Finkelstein, I.E. (1913). The marking system in theory and prac-tice. *Baltimore.* https://babel.hathitrust.org/cgi/pt?id=uc1.$b264457;view=1up;seq=9

Fleiss J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 7*, 378-382.

Fleiss, J.L. (1981) *Statistical methods for rates and proportions*. John Wiley, 38-46.

Haladyna, T.M. (1999). *A Complete guide to student grading*. Allyn and Bacon. A Viacom Company.

Han, K.T. (2011). *User's manual: SimulCAT*. http://www.umass.edu/remp/software/simcata/simulcat/SimulCAT_Manual.pdf.

Harran University (2022). Harran Üniversitesi Bağıl Değerlendirme Yönergesi [Harran University Relative Assessment Instruction]. http://ogrenci.harran.edu.tr/assets/uploads/other/files/ogrenci/files/Ba%C4%9F%C4%B1l_De%C4%9Ferlendirme_Sistemi_Ekim_2019_web.pdf

Istanbul University (2022). İstanbul Üniversitesi Önlisans, Lisans Ölçme ve Değerlendirme Esasları [Istanbul University Associate Degree, Undergraduate Measurement and Evaluation Principles]. https://cdn.istanbul.edu.tr/FileHandler2.ashx?f=olcme-degerlendirme.pdf.

Kaya, Ü., & Semerci, Ç. (2017). The opinions about relative and absolute assessment of teaching staff in the higher education. *The Journal of Academic Social Science, 5*(47), 457-467. https://doi: 10.16992/asos.12321

Kaysi, F., Bavli, B., & Gürol, A. (2017). Educational connoisseurship and criticism: evaluation of a cooperation model between university and the sector on vocational education. *Journal of Education and Practice*, *8*(6), 25-35.

Kelley, A.C., & Zarembka, P. (1968). Normalization of student test scores: An experimental justification. *The Journal of Educational Research*, *62*(4), 160-164. https://www.jstor.org/stable/27532173

Keskin M., & Ertan H. (2001). *İstanbul Üniversitesi'nin bağıl değerlendirme sistemi kitapçığı* [*Istanbul University's relative assessment system booklet*]. İstanbul: İstanbul Üniversitesi.

Kırklareli University (2022). Kırklareli Üniversitesi Sınav ve Başarı Değerlendirme Yönergesi [Kırklareli University Examination and Success Assessment Instruction]. https://oidb.klu.edu.tr/Yardimci_Sayfalar/183-yonergeler.klu

Kubiszyn, T., & Borich, G. (2000). *Educational testing and measurement*: *Classroom application and practice* (6th ed.). John Wiley & Sons, Inc.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lok, B., McNaught, C., & Young, K. (2016). Criterian-referenced and norm-reference assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education*, *41*(3), 450-465. https://doi.org/10.1080/02602938.2015.1022136

Mandernach, B.J. (2003). Effective grading strategies. *Park University Faculty Development Quick Tips*. https://www.park.edu/center-for-excellence-in-teaching-and-learning/

Martin, I.G., & Jolly, B. (2002). Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. *Medical education*, *36*(5), 418-425. https://doi.org/10.1046/j.1365-2923.2002.01207.x

Moses, M.S., & Nanna, M.J. (2007). The testing culture and the persistence of high-stakes testing reforms. *Education and Culture*, 55-72.

Nartgün, Z. (2007). Aynı puanlar üzerinden yapılan mutlak ve bağıl değerlendirme uygulamalarının notlarda farklılık oluşturup oluşturmadığına ilişkin bir inceleme [An examination of whether absolute and relative evaluation practices based on the same scores create differences in grades]. *Ege Eğitim Dergisi, 8*(1), 19-40.

Nitko, A.J., & Brookhart, S.M. (2007), *Educational assessment of students*. Pearson Education Inc.

Özçelik, D.A. (1992). *Okullarda ölçme ve değerlendirme* [*Measurement and evaluation in schools*]. ÖSYM.

Özkan, Y.Ö. (2016). Üniversite mezuniyet notları ve dönüşüm tablolarında yaşanan kaos [The chaos in university graduation grades and conversion tables]. *Journal of Higher Education and Science, 6(1)*, 71-76.

Öztürk-Gübeş, N. (2021). An investigation into weighting problem in norm-referenced grading system. *Eurasian Journal of Educational Research, 93*, 337-356. https://doi.org/10.14689/ejer.2021.93.16

Reynolds, C.R., Livingston, R.B., & Wilson, W. (2006). *The problem of bias in educational assessment in measurement and assessment in education*. Pearson Education.

Sayın, A. (2016). The Effect of using relative and absolute criteria to decide students' passing or failing a Course. *Journal of Education and Training Studies*, *4*(9), 1-9. http://dx.doi.org/10.11114/jets.v4i9.1571

Thorndike, R.M. (2005). *Measurement and evaluation in psychology and education*. Pearson Education.

Thorndike, R.L., & Hagen, E.P. (1977). *Measurement and evaluation in psychology and education*. John Wiley & Sons, Inc.

Tinkelman, D., Venuti, E., & Schain, L. (2013). Disparate methods of combining test and assignment scores into course grades. *Global Perspectives on Accounting Education, 10*, 61-80. https://gpae.wcu.edu/disparate-methods-of-combining-test-and-assignment-scores-into-course-grades/

Turgut, M.F. (1983). Program değerlendirme. *Cumhuriyet Döneminde Eğitim* [Program evaluation. Education in the Republican Era]. 215-234.

Turgut, M.F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme* [*Measurement and evaluation in education*]. Pegem Akademi.

URAP (2022). University Ranking by Academic Performance. https://newtr.urapcenter.org

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 4, 813–838

https://doi.org/10.21449/ijate.1260245

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

Global competence scale: An adaptation to measure pre-service English teachers' global competences

Ismail Emre Kos[1], Handan Celik[2*]

[1]Ahmet Uğur Turhan Science and Technology High School, Trabzon, Türkiye
[2]Trabzon University, Fatih Faculty of Education, Trabzon, Türkiye

ARTICLE HISTORY

Received: Mar. 04, 2023

Accepted: Dec. 16, 2023

Keywords:

Global competence,
Globalization,
Pre-service English teachers,
Teacher education.

Abstract: Global competence is a comprehensive term referring to the interconnectedness of various constructs ranging from knowledge to values required to communicate, cooperate, and work towards the well-being of not only the local but also the global community. Teacher education has an important role in preparing teachers equipped with global competences. Therefore, having tools which can validly and reliably measure if and to what extent pre-service teachers are globally competent is a requisite. Hence, this study aimed at adapting and validating the Global Competence scale developed by Liu et al. (2020) to Turkish to measure pre-service English language teachers' global competences and to obtain evidence regarding the psychometric properties of the scale to measure global competences in teaching and teacher education. The data collected from pre-service English language teachers (*N*=351) studying at various universities in Türkiye was divided into two equal halves. The first part of it was used to perform exploratory factor analysis which revealed an eight-factor 29-item structure. The second half which was used for confirmatory factor analysis yielded a good fit of a 25-item, eight-factor structure scale. The Cronbach's Alpha coefficient (*α*= .88) and McDonald's Omega (*ω* = .89) which indicated good internal consistency in the CFA dataset revealed excellent internal consistency (*α*= .90, *ω* = .91) in another independent dataset. Thus, the study revealed that the Global Competence scale has a good level of psychometric properties and reliability to measure pre-service English language teachers' global competences in the Turkish context.

1. INTRODUCTION

Today's world is more connected than ever before as the borders have been wiped out due to the growth in information and communication technologies and the human workforce who are linguistically more able. However, it is more challenging and demanding, as it requires individuals to possess various competences one of which is global competences. Its importance has been on the rise in the last few decades as a result of the diversity in different walks of life and the need to meet the expectations of such diversity. Moreover, global competence, as a key part of global education and a key concept in global competence education (Boix Mansilla & Jackson, 2011), is one of the noteworthy concepts for the United Nations Educational, Scientific and Cultural Organization (UNESCO) and its long-standing partner Organization for Economic

*CONTACT: Handan Celik ✉ handanelik@gmail.com Trabzon University, Fatih Faculty of Education, Trabzon, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

Co-operation and Development (OECD) which supports UNESCO so that it can achieve its goals for sustainability and education for a sustainable future. Besides enabling individuals to possess the skills to live in communities of multicultural, multilingual, and multiethnic diversity, to compete and succeed in a competitive job market, to use media and technology effectively, and to contribute to the achievement in sustainable development goals (OECD, 2018), "education for global competence can promote cultural awareness and respectful interactions in increasingly diverse societies" (p. 4). Moreover, the emphasis the United Nations (2015) and OECD (2018) put onto the education of globally competent individuals who can effectively and adequately respond to such global challenges as poverty, climate crisis, hunger, justice, and peace strengthens the need to do so. Furthermore, in his preface to Boix Mansilla and Jackson's (2011) book entitled *Education for Global competence: Preparing Our Youth to Engage the World*, Howard Gardner wrote,

> … today's students need a globally conscious education for what is assured a global era. Young people need to understand the worldwide circulation of ideas, products, fashions, media, ideologies, and human beings. These phenomena are real, powerful, ubiquitous. By the same token, growing up in the world of today -and tomorrow!- need preparation to tackle the range of pervasive problems: human conflict, climate change, poverty, the spread of disease, the control of nuclear energy. (p. x)

Due to the variety in terminology, providing a precise definition of global competence seems to challenge scholars and result in interchangeable terms (Schenker, 2019). Global competency is the most prevalent of them (Shams & George, 2006; Reimers, 2009; Li, 2013; Li & Xu, 2016; Meng et al., 2017; Baily & Holmarsdottir, 2019; Schenker, 2019; Liu et al., 2020) which is accompanied by other terms such as intercultural competency (Deardorff, 2006; Bektaş-Çetinkaya & Çelik, 2013; Cui, 2013; Polat & Barka, 2014; Lin & Kapur, 2021; van de Vijver & Leung, 2019), global awareness (Hanvey, 1982; Kirkwood, 2001; Oxfam, 2006; Merryfield, 2008; Kurt et al., 2013; Hongtao, 2013), global mindedness (Hett, 1993; Park et al., 2016), and global citizenship (Lima & Brown, 2007; Morais & Ogden, 2011; Oxfam, 2015; Başarır, 2017; Andrews & Aydin, 2020).

Despite the variety in the terminology, as comprehensively defined by Liu et al. (2020), global competence (GC hereafter) refers to "students' capabilities to actively acquire and understand other cultures and norms, keep an open mind, and use their global knowledge to communicate, interact, and work effectively outside their own culture" (p. 2). OECD (2018) defines GC as "a multidimensional capacity" and globally competent individuals as those who "can examine local, global and intercultural issues, understand and appreciate different perspectives and worldviews, interact successfully and respectfully with others, and take responsible action toward sustainability and collective well-being" (p. 4). Morais and Ogden (2011) also regard GC as a comprehensive term referring to one's self-awareness, intercultural communication, and global knowledge as part of their global citizenship. What deserves attention in all these definitions is the multidimensional nature of GC which emphasizes knowledge, understanding, communication, cooperation, and action as key tenets of GC as scholars agree upon (see Piacentini, 2017; OECD, 2018; Parmigiani et al., 2022a). UNESCO also puts a stronger emphasis on culture, education, and communication and information for sustainability and sustainable development goals for which globally competent individuals are a must. Besides, the 21st century skills movement which is largely adopted by UNESCO and OECD includes global competences as components of global citizenship (see Morais & Ogden, 2011). Through its program for international student assessment (PISA), OECD prioritizes education for global competence and assessment of global competences (see OECD, 2018; OECD & Asia Society, 2018). When they improved the curricula for grades 2-8 and 9-12, the Turkish Ministry of National Education (MoNE, 2017) also embraced the 21st-century skills framework besides values education and necessitated the injection of such global competences as communication in foreign languages, social and civic competences, cultural awareness, and communication.

Certain practices and opportunities such as study abroad (see Ozkul, 2019; Schenker, 2019; Fisher et al., 2022), exchange programs, virtual exchange programs (see Duffy et al., 2022; Ndubuisi et al., 2022), and teaching abroad (Cushner & Mahon, 2016) are seen as key means to acquire and develop GC. Such programs are also known to alter students' perspectives of the world and improve their cultural awareness (Cushner & Mahon, 2016) and intercultural competence and abilities (He et al., 2017; Özkan & Mutludoğan, 2018). Increase in social responsibility and civic mindedness and citizenship awareness (Lenkaitis & Loranc, 2019), enhanced knowledge and awareness of technological tools for communication and learning and understanding of the society are reported among other achievements (Hilliker & Loranc, 2022).

The concept of GC encompasses knowledge, understanding, and ability of local and global (Parmigiani et al., 2022a) besides language ability and knowledge and understanding of culture (Zhao, 2010). Furthermore, active engagement- that is, action - is required to influence the close and distant environments in which people live and to retain persistent consciousness in these environments in addition to attitudes and values to respond to global challenges (OECD, 2018). B. Hunter et al. (2006) regard globally competent individuals as those having "a firm understanding of the concept of globalization and world history … the recognition of the interconnectedness of society, politics, history, economics, the environment, and related topics" (p. 282). With all these in mind, Liu et al. (2020) emphasize the importance of GC for both undergraduate and graduate students as they are most likely to work and communicate in contexts of linguistic and cultural diversity. Besides, they need to perform some professional tasks such as publishing and presenting their research papers in journals or at various organizations such as international conferences or meetings among colleagues.

Furthermore, what GC implies for teacher education deserves close consideration because to meet the challenges of today's global world, "schools need teachers who understand the implications of globalization, are able to effectively work with the increasingly culturally and linguistically diverse student population, and to deliver a globally oriented curriculum" (Zhao, 2010, p. 426). For this reason, Parmigiani et al. (2022a, p. 1) regard "cooperation, inclusion, social engagement, and multicultural dialogue" as indicators of GC in teaching. Moreover, recent research on sustainability and UNESCO's sustainable development goals put paramount emphasis on the need for teacher education for a more sustainable world (see Fischer et al., 2022; Rieckmann & Barth, 2022; Rieckmann, 2023). English language teachers as those who are more likely to work with culturally and linguistically diverse student groups play a pivotal role in a broad range of issues. The cultivation of intercultural awareness, sensitivity, communication and openmindedness for the appreciation of different worldviews and cultural perspectives are some of the foremost important issues. Besides, the injection of such attitudes and values as nondiscrimination alongside with cultivation of knowledge and skills of locally and globally important issues such as poverty, climate crisis, responsible use of natural resources such as water and the awareness to take action towards a more sustainable future are some other issues. Last but not least, the cultivation of global citizenship to meet the needs of increasingly diverse language classrooms is another motivation to educate globally competent English teachers who can also raise globally competent generations. Therefore, global competence is now regarded as "a fundamental disposition for teachers" (Parmigiani et al., 2022b, p. 1).

However, looking at how studies defined GC so far, we can easily conclude that there are diverse and even too broad opinions regarding what constitutes GC which further complicates the measurement of it. Despite the immense amount of interest regarding why and how to educate globally competent teachers (see Zhao, 2010; Boix Mansilla & Jackson, 2011; Brennan

& Holliday, 2019; Tichnor-Wagner et al., 2019; Kerkhoff & Cloud, 2020) and to measure GC in teacher education programs (see Parmigiani et al., 2022a; Parmigiani et al., 2022b; Sokal & Parmigiani, 2022), due to the complexity and multifaceted nature of the concept of GC which seems to harden the development of an instrument, very few studies have so far come up with a tool. To our best knowledge, in the Turkish context, there has been no tool developed to measure neither pre-service nor in-service English language teachers' GCs. The only tool available which is adapted by Karaca Akarsu and Özdemir (2021) to measure teachers' GC is not specific to English language teachers either.

In this regard, it is crucial to validate a tool to measure pre-service English teachers' GCs in a comprehensive manner. Therefore, in the current study, we aimed to adapt the global competence scale (GSC) developed by Liu et al. (2020) to Turkish and to test its psychometric properties. In doing so, we also aimed to examine and confirm the theoretical structure of GC as suggested by Liu et al. (2020).

1.1. Tools to Measure Global Competence

Due to the complexity and multidimensional nature of the construct of GC, researchers have approached the issue from various perspectives and used various tools. In this section, so as to draw a concise and precise picture, we only present the ones that are directly aimed at measuring GC, either as a standalone construct or as a dimension of a larger, comprehensive construct such as global citizenship (see Table 1).

Table 1. *Summary of the instruments developed or adapted to measure global competence.*

Author-Year	Type of the study	Constructs-Dimensions	Instrument	Participant group	Analysis conducted
Zheldibayeva (2023)	Adaptation	Knowledge, skills, attitudes and values	Global competence scale (Liu et al., 2020)	Graduate and undergraduate educational psychology students	EFA & CFA
Parmigiani et al. (2022b)	Development	Exploring, engaging, and acting	Global competence rubric	Experts in teacher education and international/intercultural educational issues	Modified delphi method
Karaca Akarsu & Özdemir (2021)	Adaptation	Disposition, knowledge, skills	Global competence survey (Brantley Todd, 2017)	In-service teachers of various majors	EFA & CFA
Liu et al. (2020)	Development	Knowledge, skills, attitudes and values	Global competence scale	Graduate students	EFA & CFA
Brantley Todd (2017)	Development	Disposition, knowledge, skill	Global competence survey	Elementary school pre-service teachers	Delphi Technique
Şahin & Çermik (2014)	Adaptation	Social responsibility, global competence, global civic engagement	Global citizenship scale (Morais & Ogden, 2011)	Undergraduate students at various majors at Faculty of Education & Faculty of Arts and Letters	EFA & CFA

Braskamp et al. (2014)	Development	Cognitive, intrapersonal, and interpersonal	Global perspective inventory (GPI)	Undergraduate, students	EFA & Reliability
Morais & Ogden (2011)	Development	Social responsibility, global competence, global civic engagement	Global citizenship scale	Postsecondary students (undergraduate & international undergraduate programmes)	EFA & CFA
W. D. Hunter (2004)	Development	Knowledge, skills, attitude, and experiences	GC assessment instrument for evaluating college graduates	Human resource managers of transnational corporations and international educators at higher education institutions	Questionnaire Delphi technique
Hett (1993)	Development	Responsibility, cultural pluralism, efficacy, global centrism, and interconnection	Global-mindedness scale	Undergraduate students of various majors (arts, engineering, social sciences etc.)	EFA & CFA

As the summary shows, researchers' interest and work in GC in the last two decades has resulted in various, but still a limited number of, tools to measure GC. As one of the few researchers, Parmigiani et al. (2022b) developed a 32-item self-assessment rubric including 16 dimensions ranging from openness to interactive assessment strategies under three main areas named as exploring, engaging, and acting to be used by both pre-service teachers and teacher educators. Although the rubric is comprehensive, there was no statistical analysis to verify factor structure and reliability. The 48-item global competence survey developed by Brantley Todd (2017) includes three main dimensions which include the sub factors; open-mindedness, self-knowledge, communication capacity, and problem solving. The 28-item survey developed by Hunter (2004) as part of his Doctoral dissertation to evaluate the global competences of graduates of international education had four sections as knowledge, skills, attitudes, and experiences. However, the factor structure of the survey was not tested and validated through factor analyses. The 32-item global perspective inventory developed by Braskamp et al. (2014) includes 3 main dimensions named as cognitive, intrapersonal, and interpersonal. Each dimension has two scales. The 12-item cognitive dimension has cognitive knowing and cognitive knowledge scales. The 11-item intrapersonal dimension has intrapersonal identity and intrapersonal affect scales, and the 9-item interpersonal dimension has interpersonal social responsibility and interpersonal social interaction. All were tested for factor analysis and internal consistency. Additionally, Morais and Ogden (2011) developed the 30-item global citizenship scale under such factors as social responsibility, global competence, global civic engagement, self-awareness, intercultural communication, and global knowledge which they reported as strong and reliable to be used in education abroad contexts. The 30-item global-mindedness scale developed by Hett (1993) revealed a 5-factor structure namely responsibility, cultural pluralism, efficacy, global centrism, and interconnectedness through the factor analysis which also revealed acceptable levels of validity and reliability.

Additionally, a closer look into the audience that the tools can be used with revealed that although the majority included undergraduate students from various majors except for teacher education (e.g. Braskamp et al., 2014; Hett, 1993) and those of teacher education such as elementary school (e.g. Brantley Todd, 2017), Turkish language, Mathematics, Preschool teaching, Social Sciences teaching (e.g., Şahin & Çermik, 2014), none addressed undergraduate students from English language teaching programs. The global perspective inventory developed by Braskmap et al. (2014) was suggested to be potentially used in such programs as study abroad, international student orientation, service learning or with faculty members or freshmen to seniors. The global competence rubric developed by Parmigiani et al. (2022b) was suggested to be used in various contexts including before and after study or training abroad and training besides its use for self-assessment by teacher educators and pre-service teachers. Although Şahin and Çermik (2014) validated Morais and Ogden's (2011) global citizenship scale into Turkish with the participation of undergraduate students from teacher education programs, pre-service English language teachers were not among the participants. Besides, the adaptation study of the global competence survey by Karaca Akarsu and Özdemir (2021) to measure teachers' global competences did not specifically address English language teachers, despite having a small number of English language teachers in the sample. All of these instruments are, without a doubt, valuable for use in education. However, the scale developed by Liu et al. (2020) to measure the global competences of graduate students is the most comprehensive in terms of its strength in defining the theoretical structure of the concept of GC (see Table 2 for details) as revealed by factor analysis and fit indices. Besides the scale addresses such significant competences as understanding of globalization (see Altan, 2017; Block & Cameron, 2002; Gnutzmann & Intemann, 2005), cross-cultural communication (Byram, 2009; Sarıçoban & Oz, 2014), appreciation of and respect towards cultures and values (İşisağ, 2010) which apply to English language teachers, and are worth closer examination to explore if and to what extent pre-service English teachers, who have a significant role in educating globally competent individuals, are equipped with such competences. In this regard, there seems an obvious need to adapt it to Turkish. The validation of such an instrument can also encourage researchers to test its psychometric features for use with pre- and in-service English language teachers in other contexts outside of Türkiye.

2. METHOD

This study aimed to adapt the global competence scale (GSC) developed by Liu et al. (2020) to Turkish and obtain empirical evidence regarding its psychometric properties to measure pre-service English language teachers' global competences in the Turkish context. With these in mind, we employed survey methodology to collect data for validity and reliability measures. We went through the following steps as suggested by Hambleton and Patsula (1998):

- considering such factors as purpose, time, resources, expertise, and relevance of the construct across cultures and groups, we determined the tool to be validated, rather than developing one,
- contacted the authors of the original scale and requested their permission
- selected two translators for forward translation and two other translators for back translation (for more about translation and linguistic validity see section 2.2)
- employed the scale to test its psychometric properties in Turkish
- employed the scale with another group of participants to cross-check its reliability.

2.1. Instrument

The GSC was originally developed by Liu et al. (2020) as a 35-item instrument to measure graduate students' global competences. They adopted 20 items from various other tools measuring global competence (see Olson & Kroeger, 2001; Hunter et al., 2006; Li, 2013), global perspective (see Braskmap et al., 2014), and global citizenship (see Morais & Ogden, 2011) and wrote 15 items. The items were reviewed by three experts who have studied and had work experience abroad in the internationalization of higher education and graduate education. Based on their feedback and comments, Liu et al. (2020) refined the scale and pilot tested it with 68 students for clarity and relevance of items. Based on the feedback, they further refined the scale and piloted it with 1421 graduate students from five universities. Their initial exploratory factor analysis (EFA) resulted in eight factors explaining 68.6 % of the total variance. However, closer examination of item 7 revealed that it was different from all the other items. Therefore, it was reduced. Besides, item 11 was found both to be ambiguous and have strong loading in two factors (World Knowledge and Open Attitude and Values). It was also removed which resulted in a 33-item scale. The 11 items in the Attitudes and Values dimension were further analyzed through EFA revealing a three-factor structure resulting in a nine-factor model explaining 71.9 % of the total variance. The results of the confirmatory factor analysis (CFA) confirmed the nine-factor model (χ^2 (459) = 1292.5, *p*<.001, *RMSEA* = .051, *CFI* = .932, *TLI* = .922) under three dimensions (see Table 2).

Table 2. *Dimensions, sub-factors/definitions, and number of the items of the GCS* (as reported by Liu et al., 2020, p. 4).

Dimensions	Sub-factors/Definitions	Item numbers in the original scale
Knowledge and Understanding	1. World knowledge (WK): Have basic knowledge of other countries' languages, cultures, histories, and geographies	1, 2, 3
	2. Understanding Globalization (UG): Understand globalization, its developmental trends and its influence	4, 5, 6
	3. International Academic Knowledge (IAK): Have knowledge about international frontier research problems, theories, and methods	8, 9, 10
Skills	1. Use of Tools (UT): Be able to communicate in a foreign language and use information technology and other tools	12, 13, 20
	2. Cross-cultural communication (CCC): Be able to communicate, learn, and work with people from different cultural backgrounds	14, 15, 16, 17, 18, 19
	3. International Academic Communication (IAC): Be able to contact and communicate with scholars from different cultures	21, 22, 23, 24
Attitude and Values	1. Intent to Interact (II): Seek cross-cultural experiences, learning, and research.	25, 26, 27, 28, 29
	2. Open Attitude (OA): Have openness to understand, respect, and appreciate people outside one's own culture.	30, 31, 32
	3. Values (V): Identify with one's own culture and recognize that one's own worldview is not universal.	33, 34, 35

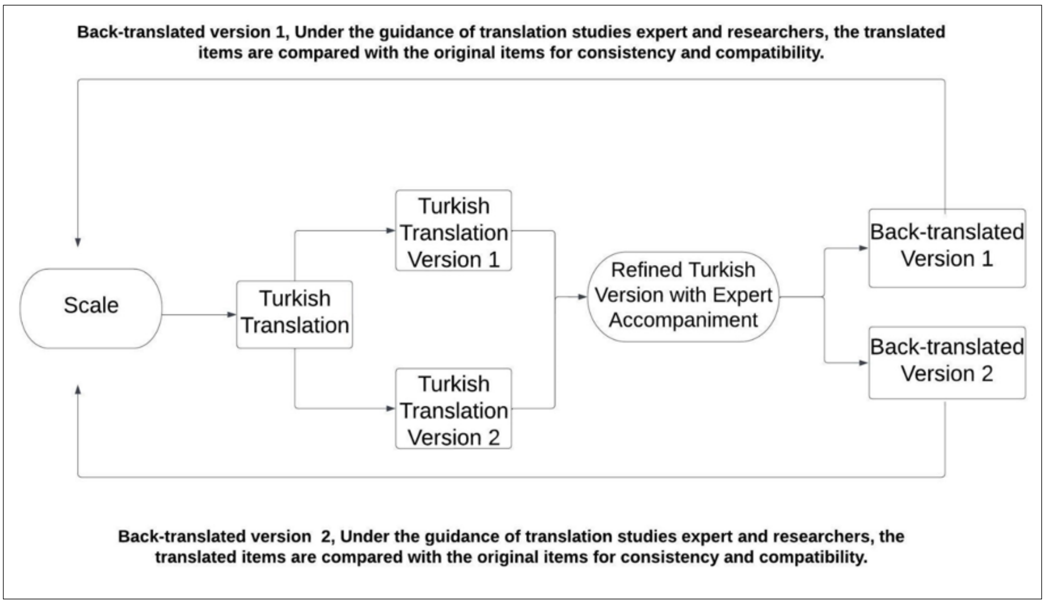
The items were designed on a five-point Likert scale ranging from strongly agree (5) to strongly disagree (1). 3 on the Likert was labeled as 'unsure'. Liu et al. (2020) emphasize the importance of GC for undergraduate students besides graduate ones. In their suggestions for further research, they state the need for translation and modification of the GCS in other countries to test its validity and reliability. Most importantly, they emphasize the need for more research using the GSC in different universities, departments, and programs.

2.2. Scale Translation

Before validating the scale, we submitted the required documents to the Research Ethics Committee of Trabzon University. Upon receiving approval from the Ethics Committee (E-

81614018-000-2200023392), we proceeded to the translation process to assure the scale's use in cross-cultural and cross-language contexts and "to achieve equivalence between the instrument in the SL [source language] and the instrument in the TL [target language]" (Sousa & Rojjanasrirat, 2011, p. 269). As we did so, we took Sousa and Rojjanasrirat's guidelines and steps into consideration for translation and adaptation (see Figure 1).

Figure 1. *Translation and backtranslation process.*


Back-translated version 1, Under the guidance of translation studies expert and researchers, the translated items are compared with the original items for consistency and compatibility.
Scale
Turkish Translation
Turkish Translation Version 1
Turkish Translation Version 2
Refined Turkish Version with Expert Accompaniment
Back-translated Version 1
Back-translated Version 2
Back-translated version 2, Under the guidance of translation studies expert and researchers, the translated items are compared with the original items for consistency and compatibility.

The items were firstly translated into Turkish by two bilingual (Turkish and English) faculty members at the English language teaching program where the second author teaches. They translated the items independently. Upon receiving the two versions of Turkish translations, we requested feedback and supervision from a third translator pursuing a Ph.D. degree and possessing teaching experience in English translation and interpretation. Using a "committee approach", as Sousa and Rojjanasrirat (2011, p. 270) regard, we had a meeting with the third translator to compare the two translated versions of the scale and to resolve ambiguities in the word choice and the syntax. This was also to eliminate any likely cultural and linguistic differences in translation which could become sources of error in adapting a scale (Hambleton & Patsula, 1998). We closely examined each item in the original scale and its translations in both Turkish versions to achieve a refined Turkish translation (a synthesis) which we sent into back-translation. For the back-translation, we worked with two other translators, one with a PhD degree in English translation and interpretation and the other with an MA degree in English language teaching. They were completely blind to the original scale and worked independently. In both Turkish translations and back-translations, we kept the translators blind, so that none of them knew who the other translator was. Receiving the back-translations, we had another meeting with the third translator who knew the refined Turkish translation very well and compared the two back-translated versions of the instrument with its original language for consistency and compatibility. After achieving consensus regarding the relevance of the back translated items to the original ones, we finalized the translation process.

2.3. Participants

The sample included pre-service English language teachers (*N*=351) studying at English language teaching programs of state and private universities (*N*=30) from different regions and cities, i.e. Erzurum to Çanakkale, in Türkiye. In this regard, the sampling was a convenience or opportunity sample addressing the participants who "meet certain practical criteria, such as geographical proximity, availability at a certain time, easy accessibility, or the willingness to volunteer" (Dörnyei, 2007, p. 99). The majority of the participants were females (*n*=234), while the rest (*n*=117) were males. Their ages ranged between 17 and 42 with an average of 21. Despite a lack of agreement (see Taherdoost et al., 2014) and various opinions regarding the sample size in scale adaptation (Osborne & Castello, 2004; Boateng et al., 2018), we aimed to achieve a sample size that satisfies the item-to-respondent ratio i.e. 5 participants for each scale item (see Büyüköztürk, 2002) and which is good enough and collects data from the right people (Osborne & Castello, 2004; Boateng et al., 2018). We collected data from July to October 2022 via Google Survey form which also included some demographic questions such as age, gender, grade, and the university of study. The first question in the form addressed voluntary participation.

In the first data set used to perform EFA (*N*= 175), the respondents were mostly females (*n*=121), while less than half were males (*n*=54). The majority (*n*=60) were 3rd grade ELT students, which was followed by the 4th (*n*=56) graders, the 2nd graders (*n*=40), and the 1st graders (*n*=19). Their ages ranged between 17 and 42 with an average of 20.9. The data set used to perform the CFA (*N*=176) also included females (*n*=113) more than males (*n*=63). However, this is quite a common situation as English language teaching programs are very well known to have female students more compared to males. Moreover, responding to the questionnaire form was completely based on true voluntariness, and we organically ended up having more female respondents. Moreover, neither Liu et al. (2020) in the original study nor we in the current adaptation study took gender as a variable. Besides, Liu et al. (2020) did have no such claim if the tool measures 'one' gender's global competences. 2nd (*n*=60) and 4th graders (*n*=52) were relatively more compared to 3rd (*n*=40 and 1st graders (*n*=24). The average age was 21.6, 17 as the youngest and 39 as the oldest.

2.4. Data Analysis

For the analysis, we split the data into two equal halves. The theoretical factor structure of the scale was examined through EFA (Field, 2018; Orçan, 2018; Costa & Sarmento, 2019) based on the relationship between the variables (items) (Büyüköztürk, 2002). When a tool is translated into another language, it does not simply mean to convey exactly the same meaning in a different language. The most important consideration also requires to asssure cultural equivalence to convey the same meaning (see van de Vijver & Tanzer, 2004) and to prevent any likely scale measurement error which results from intercultural variation (Kennedy, 2005). This is a very key consideration in scale adaptation studies to begin with EFA not only to test the accuracy of the existing factor structure but also to closely examine and reveal any likely changes in factor structure across languages and cultures (see Orçan, 2018). We used the first half of the data (*N*=175) and performed EFA in SPSS 26.0 through the Principal Axis Factoring method which does not require a normality assumption (see Costello & Osborne, 2005). Furthermore, Costello and Osborne stated that scholars agree on the use of Principal Axis Factoring as it yields the most effective possible results to apply to other samples and optimum results regarding "how many meaningful factors might be in a data set" (p. 7). Besides, we used the Promax Rotation Method as a method of Oblique Rotation to test the intercorrelations between the factors (see Ryan & Blascovich, 2015) on theoretical grounds that the factors are related but not completely independent of each other (Field, 2018).Kaiser criterion method (see Costa & Sarmento, 2019) was used, and eigenvalues were also closely examined and those above 1.0 determined the number of the factors (Costello & Osborne, 2005; Taysi & Orçan, 2020). Crossloading items were eliminated not to result in any errors in modeling, thus to improve accuracy in the number of factors (Boateng et al., 2018; Li et al., 2020).

The second half of the data (*N*=176) was used for CFA and analyzed through the Maximum Likelihood Estimation method in Jamovi 2.3.18. This was to validate the "accuracy of the structure resulting from EFA" (Orçan & Çelik, 2021, p. 1198), and thus, to confirm the theoretical structure of the scale (Costa & Sarmento, 2019). The goodness of the model fit was evaluated thoroughly via Chi-square value (χ^2), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Standardized Root Mean Square (SRMR), and Root Mean Square Error of Approximation (RMSEA) indices (Xia & Yang, 2019).

After confirming the factor structure of the scale, we tested its reliability using the second half of the data used for CFA and another independent dataset (N=150) which was gathered after adaptation. Cronbach's alpha (α) coefficients and McDonald's Omega (ω) of dimensions, factors, and items in both datasets were performed in Jamovi 2.3.18.

3. RESULTS

3.1. Exploratory Factor Analysis

Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy which revealed .849 indicated good adequacy for factor analysis procedures (Kaiser, 1970). Besides, Bartlett's test of sphericity (χ^2= 3257.042, p<.01) suggested the suitability of the data for factor analysis (Shrestha, 2021). The first EFA revealed eight factors explaining 59.895 % of the total variance. The first factor namely World knowledge had three items (items 1, 2, 3) as in the original scale. The second factor, understanding globalization had three items (items 4, 5, 6) which were also the same as the items in the original scale. Similarly, the third factor, International academic knowledge had exactly the same items (items 8, 9, 10) as in the original scale. The fourth factor, Use of tools had two items (items 12, 13). Item 20 (the number in the original scale) which was originally in the Use of tools factor was found to load onto the fifth factor, namely Cross-cultural communication which had four items (items 17, 18, 19, 20). However, some items in this factor (items 14, 15, 16 as in the original scale) were cross-loading. For instance, item 14 (.444 and .445) and item 16 (.634 and .637) loaded onto two factors with differences which were equal or less than .3 (Y. Li et al., 2020). Item 15 also loaded onto two factors where the difference was less than .17 (Costello & Osborne, 2005). The factor loads were .530 and .514. These items were also found to not load onto any of the factors in the pattern matrix. The sixth factor, International academic communication had four items (21, 22, 23, 24) as in the original scale. The seventh factor, Intent to interact, had seven items (items 25, 26, 27, 28, 29, 30, 31). However, item 30 loaded onto multiple (3) factors with factor loadings higher than .50. Regarding the loads as "sufficient[ly] strong" as Acar Güvendir and Özer Özkan (2022, p. 167) stated, we excluded it. This resulted in factor seven to have 6 items. Lastly, the eighth factor Values had 4 items (items 32, 33, 34, 35).

Therefore, we removed crossloading items (14, 15, 16, 30) from further analysis and conducted another EFA which, similar to the first EFA, revealed eight factors explaining 61.723 % of the total variance of all 29 items which Hair et al. (2010) regard as satisfactory. The KMO value of the model (.857), and Bartlett's test (χ^2= 2747.252, p <.01) indicated that the data was suitable for factor analysis. Table 3 shows the structures of eight factors and factor loadings of each item as revealed in the pattern matrix. The higher the factor loading is, the greater the contribution of the item to the related factor is (Field, 2018). Eigenvalues of each factor and the percentage of the total variance that is explained by each factor is also included. Moreover, for a better and an easier interpretation of the items, the item numbers in the original scale alongside the new item numbers are given.

Table 3. *Results of the second EFA.*

Number	Item	F1 (WK)	F2 (UG)	F3 (IAC)	F4 (UT)	F5 (CCC)	F6 (IAC)	F7 (II)	F8 (V)
I1 (*Q1)	Other than my own country, I know about the history and geography of at least one other country.	.945							
I2 (Q2)	Other than my own country, I know about the political and economic systems of at least one other country.	.845							
I3 (Q3)	Other than my own country, I know about the language, cultural norms, religions, beliefs, and customs of at least one other country.	.698							
I4 (Q4)	I understand the globalization concept and its development trends.		.826						
I5 (Q5)	I understand the effect of globalization on a country’s development, individual lifestyles and scientific research activities.		.933						
I6 (Q6)	I understand the roles of international organizations and institutions in today’s world and society.		.518						
I7 (Q8)	I know the internationally accepted theories and schools of thought in my field of study or profession.			.648					
I8 (Q9)	I know the international cutting-edge research problems, issues, and theories in my field of study or profession.			.882					
I9 (Q10)	I know the main internationally accepted research methods in my field of study or profession.			.818					
I10 (Q12)	I can easily use MS Office, PDF Reader, and other common international software.				.686				
I11 (Q13)	I can easily browse foreign language websites to obtain knowledge and the requisite information.				.817				
I12 (Q17)	I am able to quickly communicate in a common language in my interactions with people from different cultures.					.816			
I13 (Q18)	I have the ability to adjust to language and communication outside of my own culture.					.955			
I14 (Q19)	I can learn, work, and live outside of my own culture.					.698			
I15 (Q20)	I can easily comprehend foreign literature in my field of study or profession.					.467			

I16 (Q21)	When faced with problems in understanding professional literature, I can take the initiative to contact and consult the author.						.555		
I17 (Q22)	I made efforts to publish papers in SCI, SSCI, ISTP, EI, and other indexed journals or conferences with my supervisors.						.580		
I18 (Q23)	I can actively seek foreign scholars to discuss research questions and issues at international academic conferences.						.908		
I19 (Q24)	I can easily discuss research questions and issues with foreign scholars at international academic conferences.						.717		
I20 (Q25)	I would like to spend time and energy interacting with foreigners and establishing contacts.							.542	
I21 (Q26)	I would like to experience life and culture in other countries (such as through tourism).							.671	
I22 (Q27)	I would like to take the risk to experience cross-cultural learning and personal development (such as through overseas study and work).							.689	
I23 (Q28)	I would like to go abroad and experience foreign countries’ academic and research environments.							.993	
I24 (Q29)	I would like to consult foreign scholars in my areas of interest at international academic lectures and report sessions.							.738	
I25 (Q31)	When communicating with foreigners, I try to understand their cultures and values.							.539	
I26 (Q32)	When communicating with foreigners, I try to appreciate their cultures and values.								.495
I27 (Q33)	I identify with my own country’s culture and values.								.625
I28 (Q34)	I believe that my worldview is one of many equally valid worldviews.								.509
I29 (Q35)	I consider myself valuable to my country and society.								.626
Eigenvalue		9.103	2.809	2.241	1.655	1.560	1.289	1.100	1.024
Variance explained (%)		31.388	9.685	7.729	5.708	5.378	4.444	3.794	3.532

3.2. Confirmatory Factor Analysis

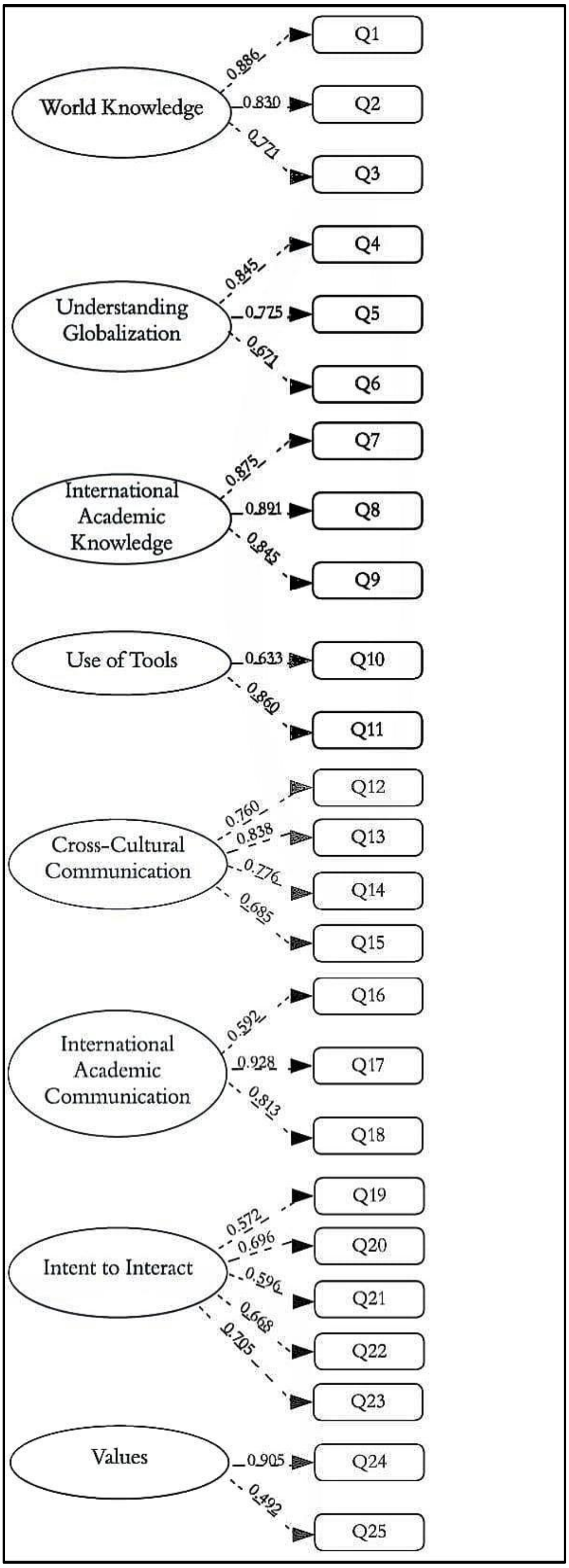
We performed the CFA of the 29 items revealed from the EFA. The initial model fit indices were as follows: χ^2= .70, *df*=349, *p*<.001, *CFI*= 0.85, *TLI*= 0.83, *SRMR*= 0.08, and *RMSEA*= 0.08. These meant that the original model needs to be reexamined and improved for a better model fit (Çapık, 2014).

A closer examination of the factor loadings of all items showed that item 33 (*I identify with my own country's culture and values.) (numbers in the original scale*) had a low loading (0.335). Therefore, we deleted the item and repeated the analysis which revealed the fit indices as χ^2= .65, *df*=322, *p*<.001, *CFI*= 0.86, *TLI*= 0.83, *SRMR*= 0.08, and *RMSEA*= 0.07. However, the indices indicated further improvement. Therefore, we also deleted item 32 (*When communicating with foreigners, I try to appreciate their cultures and values.*) which had a low load too (.443). This improved the overall goodness of fit indices χ^2= .59, *df*=296, *p*<.001, *CFI*= 0.87, *TLI*= 0.85, *SRMR*= 0.08, and *RMSEA*= 0.07, but still required further improvement. Additionally, item 21 (*When faced with problems in understanding professional literature, I can take the initiative to contact and consult the author*.) in the sixth factor and item 29 (*I would like to consult foreign scholars in my areas of interest at international academic lectures and report sessions.*) in the seventh factor were also seen not to have any relevance to the pre-service teachers as much as they did to graduate students who are more likely to attend in professional meetings and events and read research papers and consult their authors. They were also deleted which revealed the modified first-order CFA model fit indices as χ^2= .47, *df*=247, *p*<.001, *CFI*= 0.90, *TLI*= 0.87, *SRMR*= 0.06, and *RMSEA*= 0.07. In this regard, the consistent decrease in the chi-square value (Alavi et al., 2020; MacCallum et al., 1992), the increase in CFI and TLI (≥. 90) (Brown, 2015) and the decrease in SRMR (≤ .08) (Brown, 2015; Hu & Bentler, 1999) and RMSEA (≤ .05-.08) (Schermelleh-Engel et al., 2003) indices as suggested by scholars resulted in a better and a good model fit. Therefore, the structural validity of the eight-factor, 25-item scale is accepted (see Figure 2). Correlation among factors is also provided below (see Table 4).

Table 4. *Correlation matrix of the eight factors.*

Factors	1	2	3	4	5	6	7	8
1.World Knowledge								
2.Understanding Globalization	.618							
3.International Academic Knowledge	.409	.426						
4.Use of Tools	.206	.327	.312					
5.Cross-cultural Communication	.224	.361	.314	.442				
6.International Academic Communication	.483	.415	.554	.280	.411			
7.Intent to Interact	.443	.477	.155	.269	.348	.325		
8.Values	.413	.498	.330	.107	.262	.339	.372	

Figure 2. *One-order confirmatory factor analysis model.*



3.3. Reliability Analysis

Having validated the scale, we ran reliability tests in Jamovi 2.3.18. to test the internal consistency of the scale, its dimensions, and the factors. Besides Cronbach's alpha (α) which is a widespread measure of reliability, we also computed McDonald's omega (*ω)* which relies on the factor loadings in the CFA (Hayes & Coutts, 2020). As Table 5 shows, the overall reliability score of the scale yielded by Cronbach's alpha was .88, and .89 as McDonald's Omega showed. Both could be interpreted as good internal consistency (Feißt et al., 2019; Taber, 2018). The same interpretation applies to Cronbach's alpha and McDonald's Omega values of Dimension 1 and Dimension 2. As for the third dimension, although *ω* suggests acceptable consistency, α indicates moderate internal consistency (Daud et al., 2018). Additionally, to make sure if the scale can gather reliable data, we computed internal consistency scores in another independent dataset (*N*=150) which revealed a slight decrease in the scores of Dimension 1 which still indicated good reliability and an increase in the scores of Dimension 2 and 3. Overall, the scale can be interpreted as having good reliability.

Table 5. *Internal consistency scores.*

	Number of items	*Internal consistency*		*Internal consistency in the second independent dataset*	
		α	*ω*	*α*	*ω*
Entire scale	25	.88	.89	.90	.91
Dimension 1 Knowledge and Understanding	9	.88	.88	.86	.86
Factor 1 World Knowledge	3	.86	.87	.83	.84
Factor 2 Understanding Globalization	3	.81	.82	.80	.82
Factor 3 International Academic Knowledge	3	.90	.90	.87	.87
Dimension 2 Skills	9	.78	.81	.83	.84
Factor 4 Use of Tools	2	.66	.71	.72	.75
Factor 5 Cross-cultural Communication	4	.84	.84	.84	.84
Factor 6 International Academic Communication	3	.82	.83	.80	.83
Dimension 3 Attitude and Values	7	.69	.76	.82	.84
Factor 7 Intent to Interact	5	.75	.78	.85	.85
Factor 8 Values	2	.59	.62	.55	.55

4. DISCUSSION and CONCLUSION

Recent studies put strong emphasis on global competence as an "impretative" (Sinagatullin, 2019, p. 48), "a continuing challenge" (Oguro & Harbon, 2022, p. 20), and "an increasingly important disposition" (Parmigiani et al., 2022a, p. 1) in teacher education so that pre-service teachers can adequately and effectively be trained to work in classrooms with diverse students. Despite the accelerating interest in integrating GCs into teacher education (see Myers & Rivero, 2019; Chen & Lin, 2021; Diveki, 2022), a very recent study regards it as inadequate yet (see Wu & Li, 2023). Scholars developed tools (Parmigiani et al., 2022b, 2023), and few studies have assessed pre-service teachers' GCs in diverse contexts (see Parmigiani et al., 2022b; Yaccob et al., 2022). As a response to this trend in the Turkish context, researchers from other fields of teacher education have now diverted their attention to this pivotal area (see Pehlivan Yılmaz, 2023). However, to the best of our knowledge, there has been neither research developing tools nor assessing pre-service English language teachers' GCs in the Turkish context yet. Moreover, success in such recent movements as global education, global competence education, education for sustainable development, and teacher education for sustainable development require globally competent teachers so that they can raise globally competent future generations. Therefore, in the current study, we aimed to adapt the global competence scale developed by Liu et al. (2020) to Turkish and validate it to be used to measure pre-service English language teachers' GCs.

The original scale which has nine factors within 3 dimensions is revealed to have an eight-factor structure within those 3 dimensions which explained %61.723 of the total variance. Hair et al. (2010) regard this as satisfactory which means that the adapted scale can validly measure pre-service English language teachers' global competences in the Turkish context. This is also because that data involving human participants and addressing such psychological constructs as competence are almost never monodimensional and have links to other concepts (Field, 2018) as revealed in the study which verified the multidimensional nature of global competence. The dimensions, namely *knowledge and understanding, skills, and attitude and values,* were verified as in the original scale. The factor loads of the 25 items were between .49 and .93.

The factor structure of the first dimension is validated as it is in the original scale. This indicates that knowledge and understanding of GC is confirmed to have competences regarding world knowledge, understanding of globalization, and international academic knowledge which applies to pre-service English language teachers as well. Additionally, the items in this dimension could be interpreted as working well in the Turkish culture.

As for the second dimension, the current study also confirms that global competence requires use of such tools as MS Office, PDF reader as international software in addition to technological competences such as browsing foreign language websites. However, different from the original factor structure of the scale, this study showed that easy comprehension of foreign literature in one's field of study and profession is not a tool, but rather an indication of cross-cultural communication. Besides, in the current study item 14 (I can analyze and evaluate issues from the perspective of a foreign culture.), item 15 (I have made efforts to understand foreigners so that we can work or live together.), and item 16 (I can be aware of cultural differences in my interactions with people from different cultures.) were not validated as clear indicators of one's cross-cultural communication as suggested in the original scale. On the other hand, they seem to be stronger indicators of cross-cultural awareness as they indicate understanding of the home and target culture, attitudes towards culturally diverse individuals, and appreciation of cultural differences (Knutson, 2006). Moreover, these items also refer to intercultural communicative competence as they "require consideration of the ways in which people of different languages -including language learners themselves- think and act and how this might impact on successful communication and interaction" as Byram et al. (2013, p. 251) stated. Such findings also support the scholars who added intercultural communication as a dimension as they developed a tool to measure GC (Morais & Ogden, 2011). Additionally, this is in line with other scholars who approach GC from the perspective of intercultural competence (Deardorff, 2006) as it requires "*effective* and *appropriate* behaviour and communication in intercultural situations" (Deardorff, 2011, p. 66) in addition to "critical thinking …, attitudes -particularly respect (which is manifested variously in cultures), openness, and curiosity, … and the ability to see from others' perspectives" (p. 68). In this regard, further research might address if and how cross-cultural awareness, intercultural competence, and intercultural communicative competence could be added as a factor(s) to the GCS.

Moreover, item 21 and item 29 which were excluded in the CFA could easily be interpreted as resonating more with graduate students, as in the original scale, rather than they do with pre-service English language teachers who are less likely to contact the authors of research papers through various means and occasions such as e-mails or scientific meetings. Additionally, item 33 (I identify with my own country's culture and values.) which revealed the lowest factor loading in the CFA suggests that 'I identify with …' seemed to make no sense semantically in

Turkish. However, more importantly, rather than being directly linked to *values* as the factor structure in the original scale, this item suggests stronger indication of cultural identity which refers to "individual's psychological identification with a particular group" (Kim, 2007, p. 238). This interpretation also verifies empirical research which reported improvement in teachers' understandings and appreciation of cultural identities upon being trained on their global competences (see Kerkhoff & Cloud, 2020). Additionally, item 30 (When communicating with foreigners, I try to respect their cultures and values.) which was reduced in the EFA as it loaded onto multiple factors and item 32 (When communicating with foreigners, I try to appreciate their cultures and values.) which was reduced in the CFA as it had a low load and the deletion of which improved the model fit indices does never mean that respecting and appreciating cultures and values of interlocutors are not valued and important. Rather, the results could indicate that there may either be another factor which is not in the original scale or these two items are redundant as there is another item which addresses understanding culture and values (item 31). Besides, this requires a closer examination and verification of if 'respecting, understanding, and appreciating' (as the behaviours communicated through the items) differ from each other and if there is any redundancy and confusion with 'try to' as in the syntax of these three items (item 31, 32, 33).

As for the reliability of the adapted version of the GSC, based on the CFA dataset *knowledge and understanding* and *skills* dimensions were found to have good internal consistency, while the third dimension, *attitude and values* indicated moderate internal consistency (Daud et al., 2018). However, the internal consistency scores of the GSC computed in another independent dataset showed that all the three dimensions have good internal reliability (≥ .8) (Field, 2018, p. 1200), while the scale itself indicates excellent reliability (>.9) (George & Mallery, 2016, p. 240).

Consequently, EFA and CFA revealed that the adapted version of the GCS has good model fit and is valid and reliable to measure pre-service English language teachers' GCs. As the first adaptation study addressing assessment of pre-service English teachers' global competences in the Turkish context, the current study both contributes to the knowedge base of global education and global competence education and provides the Turkish teacher education community with a tool to implement and further test in their own contexts. Besides all the other issues discussed so far, one of the most important conclusions is the complexity and multifaceted nature of GC as suggested by scholars (Morais & Ogden, 2011; OECD, 2018). This is what we observed in the EFA and CFA analyses in the current study, and other researchers who adapted another global competence scale with in-service teachers (Karaca Akarsu & Özdemir, 2021). In their very comprehensive study, which tested the internal consistency of a rubric they developed to assess global competences of pre-service teachers of 10 countries, Parmigiani et al. (2023) also reported that two (engaging and acting) of the three (acting) areas which include such dimensions as global self-awareness, world views, cultural diversity, professional interaction, intercultural teaching, international practice to name a few overlapped. Therefore, as in the current study, they interpret this as an indicator of the complexity and multifacetedness of GC and suggested that while the first area could better assess the GCs of students of higher education studying at a variety of disciplines, the second and the third areas could do so to assess teacher education students' GC.

Last but not least, the scale can also be used to assess the GCs of in-service English language teachers, and it can work particularly well with English teachers pursuing a degree in graduate studies. The findings also reveal that the scale can potentially contribute to the body of knowledge on GC as a developing construct which is open to further research in the Turkish teacher education context.

4.1. Limitations, Implications and Suggestions for Further Research

To the best of our knowledge, this is the first study adapting a scale into Turkish to measure pre-service English language teachers' global competences. As revealed by psychometric properties and internal consistency scores, the adapted scale is valid and reliable. Due to the dominance of female participants, which is widely known as a characteristic of English language teacher education programs, and thus emerged as an organic factor, in the future studies, the optimization of the sample size would help reduce any gender bias.

Implications for teacher education include integration of global competence knowledge base, i.e. the knowledge, skills, attitudes, and values into the curricula regarding issues of local, global, and cultural importance besides awareness raising regarding the appreciation of worldviews, communication across cultures, and taking action for collective well-being and sustainability (see OECD & Asia Society, 2018). This also requires equipping teachers with the knowledge and skills of instructional strategies such as structured debates, organized discussions, current event discussions, playing games, project-based learning, and service learning (see OECD & Asia Society, 2018, p. 6). Besides, as scholars agree (see Liu et al., 2020; Sinagatullin, 2019), the global education movement, a part of which is global competence education is closely aligned with multicultural and intercultural education. In this regard, its goals include preparing teachers who will become adequately critical and reflective to possess such values and attitudes as tolerance, respect, recognition, and appreciation of different worldviews to effectively work with culturally and linguistically diverse students. This is a must in a rapidly changing world which urges individuals to adjust to the influx of diversity around them.

Additionally, teacher education plays an important role in the preparation of globally competent teachers who will raise future generations and cultivate a global mindset. Therefore, it requires researchers and teacher educators to be critical of the system that they are part of. This means that simply educating and preparing teachers who can and will teach the content knowledge in a particular subject area has little contribution to the societal growth. However, the more interconnected World now than ever before faces, on the other hand, some serious challenges such as climate crisis, poverty, discrimination, segregation, injustice, violence, and inequalities of gender, age, income which necessitate educating our children for empathy, tolerance, understanding, justice, human dignity, and communication so on so forth. Another reason for why we need to be concerned over if our teachers are globally competent is the significance that sustainable development carries. Teachers need to have a critical understanding, knowledge of pedagogy for global competence and education for sustainable development and ability to practice action-oriented transformative pedagogy (see Rieckmann, 2023). With all these in mind, teacher education plays a key role in the arrangement of opportunities to develop and test future teachers' global competences both in local contexts such as teaching practice in local schools and mobility programs in international contexts (see Parmigiani et al., 2022a).

Moreover, the knowledge and understanding of GC which is confirmed with all its sub-factors as world knowledge, understanding of globalization, and international academic knowledge suggests some practical implications for teacher education. Therefore, teacher education curricula should offer courses addressing knowledge and competence building in these areas. Besides, international academic communication which is revealed to correlate with such other dimensions as world knowledge and international academic knowledge could also suggest room for such innovative approaches as virtual exchange and project partnerships between higher education institutions. Lastly, the adapted scale also suggests practical implication for competence building in such areas as intercultural communication and intercultural communicative competence as indicators of global competence. Studies implementing such innovative approaches as critical sociocultural pedagogy (see Wu & Li, 2023) and sustainability education (Birdman et al., forthcoming) report improvement in global competences and intercultural communicative competence.

Acknowledgments

We owe special thanks to Dr. Özge İŞBECER for her time and feedback in translation and backtranslation processes. Special thanks also go to Dr. Mehmet KOKOÇ and Dr. Fatih ORÇAN for their feedback in statistical analysis. This research was supported by TÜBİTAK (Project No 1919B012109865).

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**: Trabzon University, 2022-6/2.19.

Authorship Contribution Statement

Ismail Emre Kos: Research design, Instrument validation, Data collection, Literature review. **Handan Celik:** Supervision, Research design, Instrument validation, Data collection, Statistical analysis, Data interpretation, Literature review, Writing and revising the manuscript.

Orcid

Ismail Emre Kos https://orcid.org/0000-0003-4219-7748
Handan Celik https://orcid.org/0000-0001-8041-6062

REFERENCES

Acar Güvendir, M., & Özer Özkan, Y. (2022). Item removal strategies conducted in exploratory factor analysis: A comparative study. *International Journal of Assessment Tools in Education*, *9*(1), 165-180. https://doi.org/10.21449/ijate.827950

Alavi, M., Visentin, D.C., Thapa, D.K., Hunt, G.E., Watson, R., & Cleary, M. (2020). Chi-square for model fit in confirmatory factor analysis. *Journal of Advanced Nursing*, *76*(9), 2209-2211. https://doi.org/10.1111/jan.14399

Altan, M.Z. (2017). Globalization, English language teaching and Turkey. *International Journal of Languages' Education and Teaching*, *5*(4), 764-776. http://dx.doi.org/10.18298/ijlet.2238

Andrews, K., & Aydin, H. (2020). Pre-service teachers' perceptions of global citizenship education in the social studies curriculum. *Journal of Social Studies Education Research*, *11*(4), 84-113. https://www.learntechlib.org/p/218549/article_218549.pdf

Baily, S., & Holmarsdottir, H.B. (2019). Fostering teacher's global competencies: Bridging utopian expectations for internationalization through exchange. *FIRE: Forum for International Research in Education*, *5*, 226-244. https://doi.org/10.30564/jiep.v4i1and2.3501

Başarır, F. (2017). Examining the perceptions of English instructors regarding the incorporation of global citizenship education into ELT. *International Journal of Languages' Education and Teaching*, *5*(4), 409-425. https://doi.org/10.18298/ijlet.2127

Bektaş-Çetinkaya, Y., & Çelik, S. (2013). Perceptions of Turkish EFL teacher candidates on their level of intercultural competence. In H. Arslan, & G. Rata (Eds.), *Multicultural education: From theory to practice*, (pp. 345-362). Cambridge Scholars Press.

Birdman, J., Çelik, H., Pandarova, I., Barron, A., Benitt, N., & Schmidt, T. (forthcoming). Connecting sustainability and culture: Building competencies through virtual exchange. In W. Leal, J. Newman, A. Lange Salvia, & L. Viera Trevisan (Eds.), *World sustainability series, North American and European perspectives on sustainability in higher education*. Springer.

Block, D., & Cameron, D. (Eds.). (2002). *Globalization and language teaching*. Routledge.

Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quinonez, H.R., & Young, S.L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Boix Mansilla, V., & Jackson, A. (2011). *Educating for global competence: Preparing our youth to engage the world*. Council of Chief State School Officers.

Brantley Todd, K. (2017). *Global competence survey development* [Doctoral Dissertation, University of Kentucky]. https://uknowledge.uky.edu/cgi/viewcontent.cgi?article=1031&context=edsc_etds

Braskmap, L.A., Braskamp, D.C., & Engberg, M.E. (2014). *Global perspective inventory (GPI): Its purpose, construction, potential uses, and psychometric characteristics*. Global Perspective Institute Inc.

Brennan, S., & Holliday, E. (2019). Preparing globally competent teachers to address P-12 students' needs: One university's story. *Global Education Review*, *6*(3), 49-64.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. (2nd ed.). The Guilford Press.

Büyüköztürk, Ş. (2002). Faktör analizi: Temel kavramlar ve ölçek geliştirmede kullanımı. [Key terms and use in scale development]. *Kuram ve Uygulamada Eğitim Yönetimi*, *32*, 470-4483.

Byram, M. (2009). Intercultural competence in foreign languages: The intercultural speaker and the pedagogy of foreign language education. In D.K. Deardorff (Ed.), *The Sage handbook of intercultural competence* (pp. 321-332). Sage.

Byram, M., Holmes, P., & Savvides, N. (2013). Intercultural communicative competence in foreign language education: Questions of theory, practice and research. *The Language Learning Journal*, *41*(3), 251-253. https://doi.org/10.1080/09571736.2013.836343

Chen, Y., & Lin, R. P. (2021). Integrating global competence into elementary school pre-service teacher education of English language in Taiwan. In A. Y. Wang (Ed.), *Competency-based teacher education for English as a foreign language* (pp. 156-167) Routledge. https://doi.org/10.4324/9781003212805

Costa, V., & Sarmento, R. (2019). Confirmatory factor analysis. A case study. arXiv:1905.05598. https://doi.org/10.48550/arXiv.1905.05598

Costello, A.B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*, Article 7 https://doi.org/10.7275/jyj1-4868

Cui, Q. (2013). *Global-mindedness and intercultural competence: A quantitative study of pre-service teachers* (Order No. 3589495) [Doctoral Dissertation, Indiana State University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/global-mindedness-intercultural-competence/docview/1430505082/se

Cushner, K., & Mahon, J. (2016). Overseas student teaching: Affecting personal, professional, and global competencies in an age of globalization. *Journal of Studies in International Education*, *6*(1), 44-58. https://doi.org/10.1177/1028315302006001004

Çapık, C. (2014). Geçerlik ve güvenirlik çalışmalarında doğrulayıcı faktör analizinin kullanımı [Use of confirmatory factor analysis in validity and reliability studies]. *Anadolu Hemşirelik ve Sağlık Bilimleri Dergisi*, *17*(3), 196-205.

Daud, K.A.M., Khidzir, N.Z., İsmail, A.R., & Abdullah, F.A. (2018). Validity and reliability of instrument to measure social media skills among small and medium entrepreneurs at Pengkalan Datu River. *International Journal of Development and Sustainability*, *7*(3), 1026-1037.

Deardorff, D.K. (2006). Identification and assessment of intercultural competence as a student outcome of internationalization. *Journal of Studies in International Education*, *10*, 241-266. https://doi.org/10.1177/1028315306287002

Deardorff, D.K. (2011). Assessing intercultural competence. *New Directions for Institutional Research*, *149*, 65-79. https://doi.org/10.1002/ir.381

Diveki, R. (2022). Global competence development in EFL teacher training – An interview study on the global content in EFL teacher trainers' courses in Hungary. *Journal of Adult Learning, Knowledge and Innovation*, *5*(2), 49-59. https://doi.org/10.1556/2059.2022.00058

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford University Press.

Duffy, L.N., Stone, G.A., Townsend, J., & Cathay, J. (2022). Rethinking curriculum internationalization: Virtual exchange as a means to attaining global competencies, developing critical thinking, and experiencing transformative learning. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, *37*(1-2), 11-25. https://doi.org/10.1080/1937156X.2020.1760749

Feißt, M., Hennigs, A., Heil, J., Moosbrugger, H., Kelava, A., Stolpner, I., Kieser, M., & Rauch, G. (2019). Refining scores based on patient reported outcomes – statistical and medical perspectives. *BMC Medical Research Methodology*, *19*(167), 1-9. https://doi.org/10.1186/s12874-019-0806-9

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.

Fischer, D., King, J., Rieckmann, M., Barth, M., Büssing, A., Hemmer, I., & Lindau-Bank, D. (2022). Teacher education for sustainable development. A review of an emerging research field. *Journal of Teacher Education*, *73*(5), 509-524. https://doi.org/10.1177/00224871221105784

Fisher, C., Hitchcock, L.I., Neyer, A., Moak, S.C., Moore, S., & Marsalis, S. (2022). Contextualizing the impact of faculty-led short-term study abroad on students' global competence: Characteristics of effective programs. *Journal of Global Awareness*, *3*(1), 1-34. https://doi.org/10.24073/jga/3/01/03

George, D., & Mallery, P. (2016). *IBM SPSS statistics 23 step by step. A simple guide and reference* (14th ed.). Routledge.

Gnutzmann, C., & Intemann, F. (Eds.). (2005). *The globalization of English and the English language classroom*. Gunter Narr Verlag.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis* (7th ed.). Pearson.

Hambleton, R.K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Socail Indicators Research*, *45*, 153-171.

Hayes, A.F., & Coutts, J.J. (2020). Use Omega rather than Cronbach's Alpha for estimating reliability. But … *Communication Methods and Measures*, *14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

He, Y., Lundgren, K., & Pynes, P. (2017). Impact of short-term study abroad program: Inservice teachers' development of intercultural competence and pedagogical beliefs. *Teaching and Teacher Education*, *66*, 147-157. https://doi.org/10.1016/j.tate.2017.04.012

Hanvey, R.G. (1982). An attainable global perspective. *Theory into practice*, *21*(3), 162-167. https://doi.org/10.1080/00405848209543001

Hett, E.J. (1993). *The development of an instrument to measure global-mindedness* [Unpublished doctoral dissertation]. University of San Diego. https://digital.sandiego.edu/dissertations/584

Hilliker, S.M., & Loranc, B. (2022). Development of 21st century skills through virtual exchange. *Teaching and Teacher Education*, Article 103646. https://doi.org/10.1016/j.tate.2022.103646

Hongtao, J.I.N.G. (2013). Global awareness: foreign language teachers' beliefs and practices. *Intercultural Communication Studies*, *22*(1), 95-116.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hunter, W.D. (2004). *Knowledge, skills, attitudes, and experiences necessary to become globally competent* [Unpublished doctoral dissertation]. Lehigh University.

Hunter, B., White, G.P., & Godbey, G.C. (2006). What does it mean to be globally competent? *Journal of Studies in International education*, *10*(3), 267-285. https://doi.org/10.1177/1028315306286930

İşisağ, K.U. (2010). The acceptance and recognition of cultural diversity in foreign language teaching. *Akademik Bakış*, *4*(7), 251-260.

Kaiser, H.F. (1970). A second generation little Jiffy. *Psychometrika*, *35*(4), 401-41. https://psycnet.apa.org/doi/10.1007/BF02291817

Karaca Akarsu, C., & Özdemir, M. (2021). Küresel yetkinlik ölçeğinin Türkçe uyarlama çalışması [Turkish adaptation study of the global competence scale]. *Uluslararası Toplum Araştırmaları Dergisi (International Journal of Society Researches)*, *18*(42), 5544-5576 https://doi.org/10.26466/opus.861584

Kennedy, D.P. (2005). Scale adaptation and ethnography. *Field Methods*, *17*(4), 412-431. https://doi.org/10.1177/1525822X05280060

Kerkhoff, S.N., & Cloud, M.E. (2020). Equipping teachers with globally competent practices: A mixed methods study on integrating global competence and teacher education. *International Journal of Educational Research*, *103*, Article 101629. https://doi.org/10.1016/j.ijer.2020.101629

Kim, Y.Y. (2007). Ideology, identity, and intercultural communication: An analysis of differing academic conceptions of cultural identity. *Journal of Intercultural Communication Research*, *36*(3), 237-253. https://doi.org/10.1080/17475750701737181

Kirkwood, T.F. (2001). Our global age requires global education: Clarifying definitional ambiguities. *The Social Studies*, *92*(1), 10-15. https://doi.org/10.1080/00377990109603969

Knutson, E. (2006). Cross-cultural awareness for second/foreign language learners. *Focus on the Classroom*, *62*(4), 591-610. https://doi.org/10.3138/cmlr.62.4.591

Kurt, M., Olitsky, N., & Geis, P. (2013). Assessing global awareness over short-term study abroad sequence: A factor analysis. *Frontiers: The interdisciplinary journal of study abroad*, *23*(1), 22-41. https://doi.org/10.36366/frontiers.v23i1.327

Lenkaitis, C.A., & Loranc, B. (2019). Facilitating global citizenship development in lingua franca virtual exchanges. *Language Teaching Research*, *25*(5), 711-728. https://doi.org/10.1177/1362168819877371

Li, J., & Xu, J. (2016). Investigating causality between global experience and global competency for undergraduates in contemporary China's higher education: A transformative learning theory perspective. *International Journal of Higher Education*, *5*(3), 155-167. https://doi.org/10.5430/ijhe.v5n3p155

Li, Y. (2013). Cultivating student global competence: A pilot experimental study. *Decision Sciences Journal of Innovative Education*, *11*(1), 125-143. https://doi.org/10.1111/j.1540-4609.2012.00371.x

Li, Y., Wen, Z., Hau, K.T., Yuan, K.H., & Peng, Y. (2020). Effects of cross-loadings on determining the number of factors to retain. *Structural Equation Modelling: A Multidisciplinary Journal*, *27*(6), 841-863. https://doi.org/10.1080/10705511.2020.1745075

Lima, C.O., & Brown, S.W. (2007). Global citizenship and new literacies providing new ways for social inclusion. *Psicologia Escolar e Educacional*, *11*, 13-20. https://doi.org/10.1590/S1413-85572007000100002

Lin, C.C., & Kapur, K. (2021). Pre-service teachers' perception toward global learning experiences: implications for teacher intercultural competency development. *Journal of Narrative and Language Studies*, *9*(17), 257-270.

Liu, Y., Yin, Y., & Wu, R. (2020). Measuring graduate students' global competence: Instrument development and an empirical study with a Chinese sample. *Studies in Educational Evaluation*, *67*. https://doi.org/10.1016/j.stueduc.2020.100915

MacCallum, R.C., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490-504. https://doi.org/10.1037/0033-2909.111.3.490

Meng, Q., Zhu, C., & Cao, C. (2017). The role of intergroup contact and acculturation strategies in developing Chinese international students' global competence. *Journal of Intercultural Communication Research*, *46*(3), 210-226. https://doi.org/10.1080/17475759.2017.1308423

Merryfield, M.M. (2008). Scaffolding social studies for global awareness. *Social Education*, *72*(7), 363-366.

Morais, D.B., & Ogden, A.C. (2011). Initial development and validation of the global citizenship scale. *Journal of Studies in International Education*, *15*(5), 445- 466. https://doi.org/10.1177/1028315310375308

Ministry of National Education. (MoNE). (2017). *Müfredatta yenileme ve değişiklik çalışmalarımız üzerine*. MoNE.

Ndubuisi, A., Marzi, E., Mohammed, D., Edun, O., Asare, P., & Slotta, J. (2022). Developing global competence in global virtual team projects: A qualitative exploration of Engineering students' experiences. *Journal of Studies in International Education*, *26*(2), 259-278. https://doi.org/10.1177/10283153221091623

Myers, J.P., & Rivero, K. (2019). Preparing globally competent preservice teachers: The development of content knowledge, disciplinary skills, and instructional design. *Teaching and Teacher Education*, *77*, 214-229. https://doi.org/10.1016/j.tate.2018.10.008

OECD. (2018). *Preparing our youth for an inclusive and sustainable world: The OECD PISA global competence framework*. OECD. https://www.oecd.org/education/Global-competency-for-an-inclusive-world.pdf

OECD, & Asia Society. (2018). *Teaching for global competence in a rapidly changing world*. OECD. Available at https://asiasociety.org/sites/default/files/inline-files/teaching-for-global-competence-in-a-rapidly-changing-world-edu.pdf

Oguro, S., & Harbon, L. (2022). Enchancing pre-service teachers' global competencies through interdisciplinary study abroad. In M.D. Ramirez-Verduo, & B. Otcu-Grillman (Eds.), *Interdisciplinary approaches toward enchancing teacher education* (pp. 20-32). IGI Global.

Olson, C.L., & Kroeger, K.R. (2001). Global competency and intercultural sensitivity. *Journal of Studies in International Education*, *5*(2), 116-137. https://doi.org/10.1177/102831530152003

Orçan, F. (2018). Exploratory and confirmatory factor analysis: Which one to use first? *Journal of Measurement and Evaluation in Education and Psychology*, *9*(4), 414-421. https://doi.org/10.21031/epod.394323

Osborne, J.W., & Castello, A.B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research, and Evaluation*, *9*(9), 1-9. https://doi.org/10.7275/ktzq-jq66

Ozkul, P. (2019). *Assessing study abroad relationship with perceived global competence levels of undergraduate business students* [Doctoral dissertation, University of South Florida]. USF Tampa Graduate Theses and Dissertations. https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=9595&context=etd

Oxfam. (2015). Education for global citizenship: A guide for schools. Oxfam Education and Youth. Online. https://oxfamilibrary.openrepository.com/bitstream/handle/10546/620105/edu-global-citizenship-schools-guide-091115-en.pdf?sequence=11&isAllowed=y

Oxfam. (2006). Education for global citizenship: A guide for schools. http://www.oxfam.org.uk/coolplanet/teachers/globciti/wholeschool/getstarted.htm

Özkan, A., & Mutdoğan, S. (2018). Erasmus programı'nın tasarım öğrencilerinin yaşam ve eğitimlerine katkısı [Contribution of Erasmus Programme for Design Students' Life and Education]. *Turkish Online Journal of Design Art and Communication*, *8*(2), 153-165. https://dergipark.org.tr/tr/pub/tojdac/issue/36245/408257

Park, H.S., Durkee, C., & Slobuski, T. (2016). *Global mindedness and global citizenship education*. Oxford University Press.

Parmigiani, D., Jones, S.L., Kunnari, I., & Nicchia, E. (2022a). Global competence and teacher education programmes. A European perspective. *Cogent Education, 9*(1). https://doi.org/10.1080/2331186X.2021.2022996

Parmigiani, D., Jones, S.L., Silvaggio, C., Nicchia, E., Ambrosini, A., Pario, M., Pedevilla, A., & Sardi, I. (2022b). Assessing global competence within teacher education programmes. How to design and create a set of rubrics with a modified delphi method. *SAGE Open*, *12*(4), 1-13. https://doi.org/10.1177/21582440221128794

Parmigiani, D., Nir, A.B., Ferguson-Patrick, K., Baruch, A.F., Heddy, E., Impedovo, M.A., Ingersoll, M., Jones, M., Kimhi, Y., Lourenço, M., Macqueen, S., Pennazio, V., Sokal, L., Timkova, R., Westa, S., & Wikan, G. (2023). Assessing the development of global competence in teacher education programmes: Internal consistency and reliability of a set of rubrics. *Higher Education Pedagogies*, *8*(1), Article 2216190. https://doi.org/10.1080/23752696.2023.2216190

Pehlivan Yılmaz, A. (2023). Determination of global competence levels of pre-service social studies teachers. *e-International Journal of Educational Research*, *14*(3), 267-282. https://doi.org/10.19160/e-ijer.1233534

Piacentini, M. (2017). Developing an international assessment of global competence. *Childhood Education*, *93*(6), 507-510. https://doi.org/10.1080/00094056.2017.1398564

Polat, S., & Ogay Barka, T. (2014). Preservice teachers' intercultural competence: A comparative study of teachers in Switzerland and Turkey. *Eurasian Journal of Educational Research*, 54, 19-38. http://dx.doi.org/10.14689/ejer.2014.54.2

Reimers, F. (2009). Global competency is imperative for global success. *Chronicle of Higher Education*, *55*(21), A29. Chronicle.

Rieckmann, M., & Barth, M. (2022). Educators' competence frameworks in education for sustaibale development. In P. Vare, N. Lausselet, & M. Rieckmann (Eds.), *Competences in education for sustaible development. Critical perspectives* (pp. 19-26). Springer.

Rieckmann, M. (2023, August 31). *Education for sustainable development in teacher education and schools – Transformative and disruptive education* [Conference presentation]. Summer School Teaching for Sustainable Development, Marburg, Germany.

Ryan, W.S., & Blascovich, J. (2015). Measures of attitudes towards sexual orientation: Heterosexism, homophobia, and internalized stigma. In G.J. Boyle, D.H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 719-751). Elsevier. https://doi.org/10.1016/B978-0-12-386915-9.00025-5

Sarıçoban, A., & Oz, H. (2014). Research into pre-service English teachers' intercultural communicative competence (ICC) in Turkish Context. *Anthropologist*, *18*(2), 523-531. http://dx.doi.org/10.1080/09720073.2014.11891570

Sinagatullin, I.M. (2019). Developing preservice elementary teachers' global competence. *International Journal of Educational Reform*, *28*(1), 48-62. https://doi.org/10.1177/1056787918824193

Şahin, İ.F., & Çermik, F. (2014). Küresel vatandaşlık ölçeğinin Türkçeye uyarlanması: Güvenirlik ve geçerlik çalışması [Turkish adaptation of global citizenship scale: Reliability and validity]. *Doğu Coğrafya Dergisi*, *19*(31), 207-218. https://doi.org/10.17295/dcd.30443

Schenker, T. (2019). Fostering global competence through short-term study abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *31*(2), 139-157. https://doi.org/10.36366/frontiers.v31i2.459

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.

Shams, A., & George, C. (2006). Global competency: An interdisciplinary approach. *Academic Exchange Quarterly*, *10*(4), 249-257.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4-11. https://doi.org/10.12691/ajams-9-1-2

Sokal, L., & Parmigiani, D. (2022). Global competence in Canadian teacher candidates. *Frontiers in Education*, *7*, Article 939232. https://doi.org/10.3389/feduc.2022.939232

Sousa, V.D., & Rojjanasrirat, W. (2011). Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: A clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice*, *17*(2), 268-274. https://doi.org/10.1111/j.1365-2753.2010.01434.x

Taber, K.S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48,* 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Taherdoost, H., Sahibuddin, S., & Jalaliyoon, N. (2014). Exploratory factor analysis: Concepts and theory. In J. Balicki (Ed.), *Advances in applied and pure Mathematics* (pp. 375-382). HAL Science Ouverte. Online https://hal.science/hal-02557344

Taysi, E., & Orçan, F. (2020). Kendinden tiksinme ölçeğinin psikmetrik niteliğinin tespit edilmesi [Testing psychometric properties of the self-disgust scale]. *SDÜ Fen-Edebiyat Fakültesi Sosyal Bilimler Fakültesi Dergisi*, *50*, 167-176.

Tichnor-Wagner, A., Parkhouse, H., Glazier, J., & Cain, J.M. (2019). *Becoming a globally competent teacher*. Hawker Brownlow Education.

United Nations (UN). (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations.

van de Vijver, F., & Tanzer, N.K. (2004). Bias and equilavance in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119-135. https://doi.org/10.1016/j.erap.2003.12.004

van de Vijver, F., & Leung, K. (2009). Methodological issues and researching intercultural competence. In D.K. Deardorff (Ed.), *The Sage handbook of intercultural competence* (pp. 404-419). Sage Publications.

Wu, X., & Li, J. (2023). Becoming competent global educators: Pre-service teachers' global engagment and critical examination of human capital discourse in glocalized contexts. *International Journal of Educational Research*, *119*, Article 102181. https://doi.org/10.1016/j.ijer.2023.102181

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modelling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*, 409-428. https://doi.org/10.3758/s13428-018-1055-2

Yaccob, N.S., Yunus, M., & Hashim, H. (2022). Globally competent teachers. English as a second language teachers' perceptions on global competence in English lessons. *Frontiers in Psychology*, 13. https://doi.org/10.3389/fpsyg.2022.925160

Zhao, Y. (2010). Preparing globally competent teachers: A new imperative for teacher education. *Journal of Teacher Education*, *61*(5), 422-431. https://doi.org/10.1177/0022487110375802

Zheldibayeva, R. (2023). The adaptation and validation of the global competence scale among educational psychology students. *International Journal of Education and Practice*, *11*(1), 35-46. https://doi.org/10.18488/61.v11i1.3253