e-ISSN: 2148-7456

a peer-reviewed online journal

hosted by DergiPark

International Journal of Assessment Tools in Education

Volume: 10 Issue: 1 March 2023

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

e-ISSN 2148-7456

https://dergipark.org.tr/en/pub/ijate
https://www.ijate.net

Volume 10 Issue 1 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 Info : Dr. Izzet KARA
Address : Pamukkale University, Faculty of Education,
Kinikli Yerleskesi, 20070, Denizli, Türkiye
Phone : +90 258 296 1036
Fax : +90 258 296 1200
E-mail : ikara@pau.edu.tr

Frequency : 4 issues per year (March, June, September, December)
Online ISSN : 2148-7456
Website : http://www.ijate.net/
http://dergipark.org.tr/en/pub/ijate

Journal Contact : Dr. Eren Can AYBEK
Address : Department of Educational Sciences, Pamukkale University,
Faculty of Education, Kinikli Yerleskesi, Denizli, 20070, Türkiye
E-mail : erencanaybek@gmail.com
Phone : +90 258 296 1050

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

Review Articles

International Journal of Assessment Tools in Education

2023, Vol. 10, No. 1, 1–11

https://doi.org/10.21449/ijate.1212539

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

The bibliometric journey of IJATE from local to global

Orhan Karamustafaoglu[1,*], Metin Orbay[1], Izzet Kara[2]

[1]Amasya University, Faculty of Education, Department of Maths and Science Education, Amasya, Türkiye.
[2]Pamukkale University, Faculty of Education, Department of Maths and Science Education, Denizli, Türkiye.

ARTICLE HISTORY

Received: Nov. 30, 2022

Accepted: Feb. 22, 2023

Keywords:

Bibliometric,
Citation analysis,
IJATE,
Usage count,
Web of Science.

Abstract: *International Journal of Assessment Tools in Education* (IJATE) is one of the educational journals that is indexed in major worldwide databases such as Web of Science (WoS) and ERIC. This study presents the bibliometric characteristics of articles published in IJATE between 2014 and 2021 through the bibliometric analyses. Harzing's "Publish or Perish software" was used to collect citation data from WoS and Google Scholar databases as a tool to analyze the impact of articles. Firstly, when contributing institutions are analyzed, especially in recent, it is seen that researchers from countries such as France and Kuwait have been contributing to the journal with publications produced through international collaboration. Moreover, when the average citation numbers per article is calculated, it is understood that Australia (13) and Canada (3.5) are the countries that contribute significantly to the visibility of the journal. Such a trend will contribute significantly to the international recognition of the journal soon. On the other hand, there is a statistically significant positive relationship ($r=0.339; p<0.01$) between usage count and the number of citations by WoS. Our results reveal that while the number of references used in the articles was in consistent with the literature, the average article title lengths (12±3) were slightly longer than the ideal length (10±3). The results will provide important contributions to editors, reviewers, and authors in the journey of IJATE from local to global. The findings can guide authors, the editors and referees and also serve as a potential roadmap for the future studies and journal.

1. INTRODUCTION

Scholarly journals play a crucial role as one of sciences official communication languages in the process of revealing, disseminating and using knowledge (Hicks, 2012). Nowadays, they have become more prominent as a widely used communication tool due to reasons such as internationalization in the field of education, rapid development of the field's relations with other disciplines over time and developments in information technologies (Aman & Botte 2017; Aktaş & Karamustafaoğlu, 2022; Budd & Magnuson, 2010; Goodyear et al. 2009; Orbay et al. 2021).

Developments in information technologies have made it possible to access information easily and economically, and the information that can be accessed has increased exponentially over time (Fire & Guestrin, 2019). The competitive environment created by the increasing number

*CONTACT: Orhan KARAMUSTAFAOGLU ✉ orseka@yahoo.com 🖃 Amasya University, Faculty of Education, Department of Maths and Science Education, Amasya, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

of journals has brought along the idea of "publish or perish!" among researchers and also questions of "quality or quantity?" in conducted studies (Civera et al., 2020; Van Dalen, 2021). Therefore, understanding the characteristics of journals and publications, tracking publications, and analyzing journals are essential for understanding the present and making inferences about the past and the future. In this context, analyzing the journals with various mathematical and statistical techniques is widely used as a quality measurement tool (Donthu et al., 2021). Bibliometric analysis is the leading methods used to measure the contribution of journals to the field of science (Pritchard, 1969).

Web of Science (WoS), The Educational Resource Information Center (ERIC) and Scopus are the main databases used for bibliometric analysis of publications published in journals accepted in the field of education in academic platforms (Pranckutė, 2021). On the other hand, in the field of education, the social conditions, political and cultural climate of the region in which the country lives are taken into consideration (Özenç Uçak & Al, 2008). Therefore, journals published by countries in the field of education are evaluated in some national databases created using various criteria. The TRIndex, which is also used in academic promotions in Turkey, is one example of a database that includes national journals in the field of education (HUIC, 2022).

“International Journal of Assessment Tools in Education (IJATE)- e-ISSN: 2148-7456”, which started in 2014 as open access in the educational research field in Turkish and English languages, is one of the journals that contributed to the field by being indexed in many national and international indexes (IJATE, 2022). IJATE started its academic publishing life as two issues a year, and since 2018, it has been published as four issues a year and only in English. Researchers are not charged for article evaluation or publication processes in the journal. Studies submitted to the journal are subjected to double-blind peer-review and peer evaluation. The main purpose of the journal is to bring together the studies focused on measurement and evaluation carried out at all levels of education with the relevant stakeholders. Since 2014, IJATE has been published on the DergiPark platform, which provides electronic hosting and editorial process management services for academic refereed journals (ULAKBIM, 2022). In this context, it has come a long way from local to global by being indexed in many national and international indexes such as TRIndex in 2016, Emerging Sources Citation Index (ESCI) under WoS in 2017, and ERIC in 2018.

In this study, the articles published in IJATE between 2014 and 2021 were examined by the use of bibliometric analysis and answers were sought to the following research questions (RQ):

RQ1: How do the article numbers and authors change over the years? Which institutions are the most productive? What is the international contribution to the journal and collaboration between countries?

RQ2: Is there any relationship between the usage count of articles in WoS and the citation numbers received from WoS and Google Scholar (GS) databases?

RQ3: Which keywords are commonly used in the articles? Are they compatible with the objectives of the journal?

RQ4: What are the journal quartiles of the most cited journals in the journal and the journals that cite the journal the most?

RQ5: How does the number of references used in the articles change over the years? Are the title lengths consistent with international literature?

2. METHOD

In this study was used bibliometric analysis technique. There are a total of 247 documents published in IJATE and indexed in WoS from 2014 to 2021, including 233 articles, 8 editorial materials and 6 reviews. Of these documents, editorial materials were not included in the evaluation. In the evaluation process carried out by means of the bibliometric analysis, 239 studies constituted the study sample. These studies will be referred to as "article" hereafter. The tag information and content analysis of the articles were carried out with the data obtained from the WoS database.

As is well known, the "*Publish and Perish software*" can be used as an analysis instrument of the impact of the research by analyzing the citations (Harzing, 2007). Thus, it was used to determine the citation numbers. In this software, WoS and GS databases were selected in accordance with the purpose of the research. Repeated citations to the relevant articles in GS databases were removed. The citation search was conducted between 01-10 October 2022.

The significance level for statistical tests was accepted as $p<0.05$ and IBM SPSS Statistics for Windows, Version 26.0, was used to analyze the data. For the analysis of collected data and illustration of the bibliometric maps of scientific relations, VOSviewer 1.6.13 was used (Van Eck & Waltman, 2010).

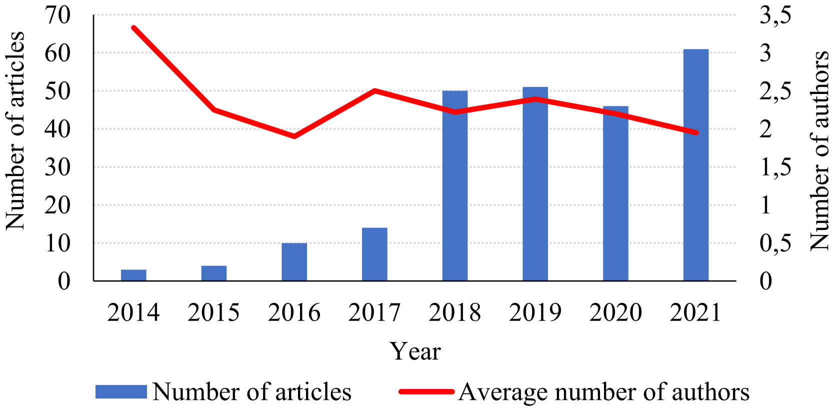
3. RESULT

The findings acquired for the study topics are provided in this part, and the interpretations are explored in the presence of relevant literature.

3.1. Results and Discussion for RQ1

The changing number of articles and authors published in IJATE by years was indicated in Figure 1. As seen in Figure 1, 239 articles were published between the years studied and the number of articles tends to increase over the years. While IJATE published a single issue in 2014 and two issues until 2017, it started to publish 4 issues in the following years. The main reason for this trend can be explained by the new start of the journal and the limited number of articles published in the first years. On the other hand, although the average number of authors per article seems to be decreasing, it was found that the median author values for all years was two (median=2), except the first year. Henriksen (2016) found that publications in educational research were written by one author during 1980-2000 and by two authors between 2001 and 2013. Similarly, the median value for authors was found to be two in the data of a journal publishing in the education field during the years 2014-2021 (Orbay et al., 2021). Therefore, the finding obtained in this study in terms of the number of authors is in harmony with the studies in the international literature.

Figure 1. *Annual distribution of articles and authors between 2014 and 2021.*



The start of the journal to be indexed in internationally recognized indexes in the education field such as ESCI in 2017 and ERIC in 2018 can be interpreted as an increasing interest in the journal. The interest brings about a positive correlation between quality and quantity of articles. One reason for this is that publishing in high impact journals that are indexed in major indexes is regarded prestigious in academia, and it is a natural byproduct of supply and demand (Arslan et al., 2022; Huang, 2016).

A Total of 414 citations from WoS were made to the studies published in the journal between 2014 and 2021. In other words, the average number of citations per article was 1.73. However, 44.76% of these articles were not cited from WoS and 20.08% were not cited from GS. It is expected that these articles would start to receive a certain citation number in the future due to the nature of educational sciences, as the studies published in the last two years were intense (71.96% for WoS, 81.25% for GS). The number of articles, total citations and total link strength, which is seen as a measure of inter-institutional collaboration, of the ten most productive institutions were given in Table 1. It is seen that all these institutions are addressed in Türkiye. Among these institutions, articles from Anadolu and Akdeniz Universities were cited well above the average, while articles from Gazi and Ankara Universities were cited below the average of the journal.

Table 1. *The first ten institutions by total articles during 2014-2021.*

No	Institution	Number of Articles	Citations
1	Hacettepe University	34	51
2	Pamukkale University	30	69
3	Gazi University	15	4
4	Ministry of National Education	12	9
5	Ankara University	12	7
6	Anadolu University	7	49
7	Akdeniz University	5	24
8	Kilis 7 Aralık University	5	7
9	Çukurova University	5	6
10	Abant Izzet Baysal University	5	5

If the content of the journal is targeted at an international audience, it would be desirable to have an international diversity of authors who can contribute to this goal. However, when the top ten contributing institutions are analyzed, IJATE's definition of "*Globally national-locally international journal*" is evoked (Pajić & Jevremov, 2014). This definition was introduced to the journals that have few international authors or readers and relatively do not receive citations from articles published in international journals (Pajić & Jevremov, 2014). It is known that this was the case for journals with Turkey addresses in citation indexes before such a definition was introduced to the literature (Doğan, Dhyi & Al, 2018; Tonta, 2017).

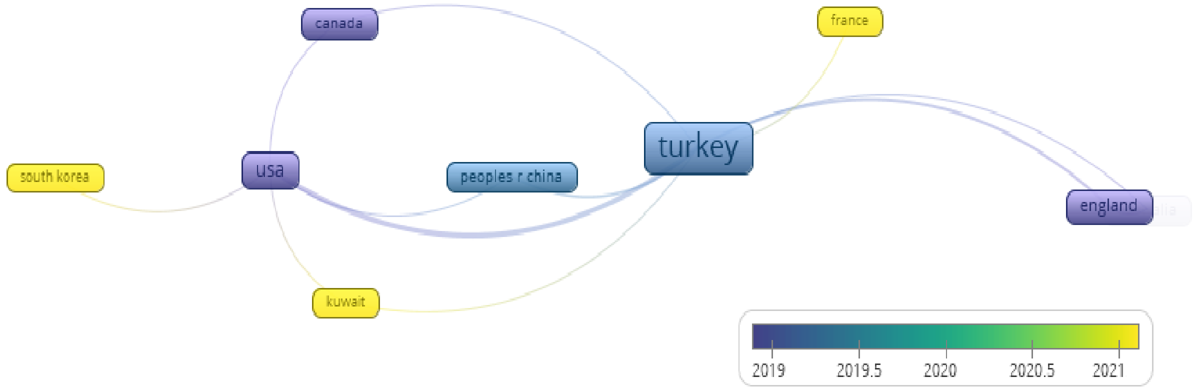
Researchers from 28 different countries have contributed to the journal so far. The article numbers, total citation numbers and Total Link Strength (TLS) values of the top ten contributing countries are given in Table 2. On the other hand, Figure 2 indicates the collaboration network of the contributing countries.

Table 2. *The most ten productive countries based on the number of articles.*

No	Country	Number of Articles	Citations	Total Link Strength
1	Turkey	190	308	17
2	USA	27	39	12
3	England	10	32	4
4	Iran	5	7	2
5	China	3	9	3
6	Australia	2	26	3
7	Canada	2	7	2
8	Ghana	2	4	0
9	Kuwait	2	2	2
10	Saudi Arabia	2	1	2

As can be seen from Table 2, when the average citation numbers per article is calculated, it is understood that Australia (13), Canada (3.5) and the UK (3.2) are the countries that contribute significantly to the visibility of the journal. The collaboration network between the contributing countries is displayed in Figure 2. Especially in recent years, it is seen that researchers from countries such as France, South Korea and Kuwait have been contributing to the journal with publications produced through international collaboration. Such a trend will contribute significantly to the international recognition of the journal soon.

Figure 2. *Overlay visualization map for the collaboration among countries.*



3.2. Results and Discussion for RQ2

The number of "Usage Count (UC)" is used as a level of interest shown by researchers in an article indexed in the WoS database (Wang et al., 2016). This criterion shows how many times the article has been read from the publisher's website directly or via the open URL, or how many times the article has been saved for use in the researcher's library. In this study, the correlation coefficient was calculated to reveal the relationship between UC and WoS and GS citations of each article. Before the analysis, the descriptive statistics results were calculated for all three data sets, and it was seen that the data did not show a normal distribution. Therefore, the Spearman correlation value between UC and the citations received by the article was calculated and indicated in Table 3.

Table 3. *Spearman Correlation Matrix among some bibliometric indicators.*

Bibliometric indicators	A	B	C
A Usage Count	1	*0.339**	*0.378**
B WoS Citation		1	*0.754**
C GS Citation			1

*Significantly correlated when the significance level is set at 0.01 (two-tailed).

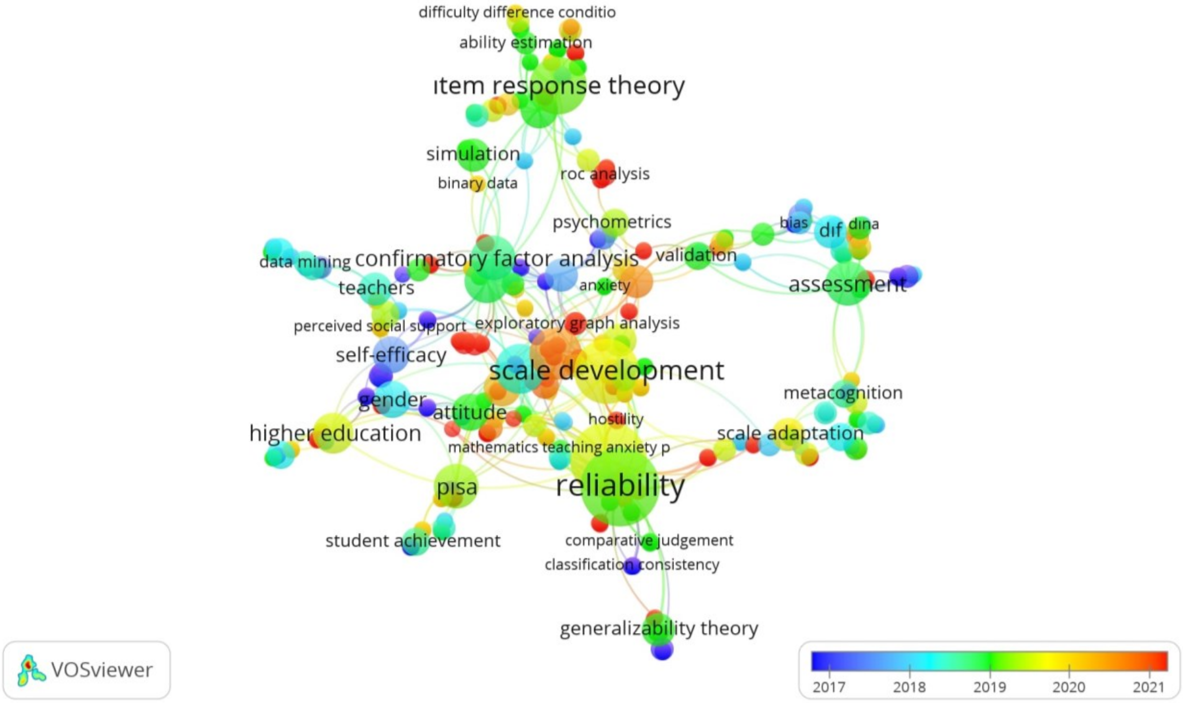
Table 3 shows that there is a statistically significant positive correlation (*r=0.339*) between UC and the number of citations by databases. In a similar vein, Nemati-Anaraki et al. (2019) found the relationship between UC and WoS citations at *r=0.401*. Furthermore, from the data obtained, a highly significant positive relationship (*r=0.754*) between WoS and GS citations is noteworthy. Martin-Martin et al. (2018) pointed out that the scope of the GS database is very wide, including the WoS database (95%), and found that almost half of the citations are from documents outside the journal. They emphasized that most of these citations were not in English. On the other hand, they found a very strong correlation of 0.92 between citations in WoS and GS databases when the category of educational research was considered.

3.3. Results and Discussion for RQ3

A total of 803 keywords were used published articles in the journal. The number of keywords used more than once is 85, meaning about 90% of the keywords were used only once. It is possible to see similar findings in the related literature (Chuang et al. 2007; Dong et al., 2012). Three main reasons for this case are taken into account as follows; *i)* the target audience of the journal is wide, but the number of articles is still limited, *ii)* some of the keywords suggested in the articles are very general concepts (design, proof, error, etc.), and *iii)* some of the keywords in the articles are acronyms or abbreviations specific to that article (MIMIC, PPSE P10, RSS, etc.).

An overlay visualization map for all keywords used in the articles is indicated in Figure 3. The circle sizes in the figure represent the frequency of use of the keywords. The five most frequently (*f*) used keywords are reliability (*f=22*), validity (*f=21*), scale development (*f=15*), item response theory (*f=12*) and measurement invariance (*f=10*). It can be stated that each keyword that emerged is fully compatible with the journal.

Figure 3. *Overlay visualization map of relationship among the most frequently used keywords.*



3.4. Results and Discussion for RQ4

The ten most cited and citing journals in IJATE with their active journal quartiles in WoS are indicated in the Table 4. As journals can be included in more than one category within WoS, optimistic mode was used in journal quartiles (Liu, Hu & Gu, 2016; Orbay et al., 2020). Within the coverage of WoS, a total of 414 citations were made to the journal, and it is seen that these were cited from high impact journals (Q1 & Q2). At the meantime, when the total number of citations is considered, it is understood that it has been cited by many different journals. This can be interpreted as the first sign of the recognition of the journal in the international literature. On the other hand, when the sources used by the articles published in the journal within the scope of WoS are examined, it is seen that high impact journals unique to the field are cited.

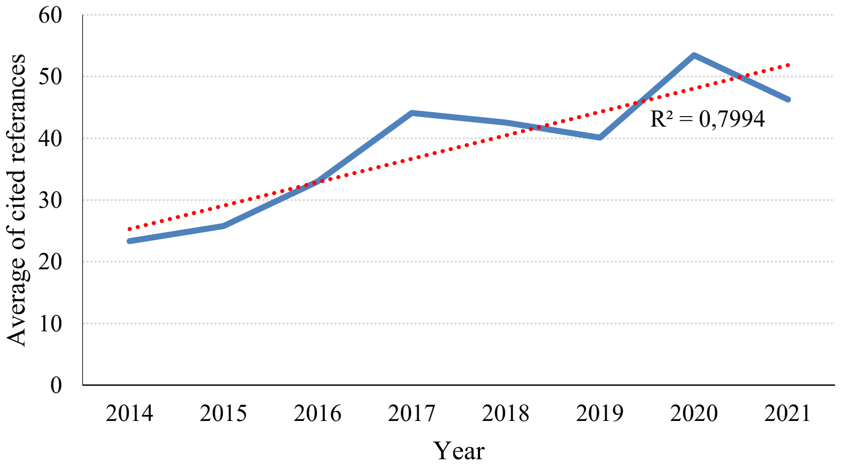
Table 4. *The most cited and citing journals in IJATE and their active journal quartiles in WoS.*

	Cited Journal in WoS			Citing Journal in WoS		
No	Journal	Active Quartile	Total Cited	Journal	Active Quartile	Total Citing
1	Front Psychol	Q1	7	App Psych Meas	Q3	70
2	Int J Assess Tools E	ESCI	7	Struct Equ Modeling	Q1	70
3	Sustainability	Q2	6	Educ Psychol Meas	Q1	54
4	Think Skills Create	Q1	3	Psychometrika	Q2	38
5	Educ Inf Technol	Q1	2	J Educ Meas	Q3	37
6	Educ Sci	ESCI	2	App Meas Educ	Q3	28
7	Eurasian J Educ Res	ESCI	2	Psychol Methods	Q1	26
8	IEEE Access	Q2	2	Egit Bilim	Q4	24
9	Nurs Educ Today	Q1	2	Pers Indiv Differ	Q2	18
10	Res Pap Educ	Q3	2	J Meas Eval Educ Psy	ESCI	17

3.5. Results and Discussion for RQ5

The change in the number of cited references in the issued articles in the journal through 2014-2021 is indicated in Figure 4. The average cited reference numbers started from 23.33 in 2014 and reached 44.24 in the period studied. Moreover, it was observed that the number of cited references tended to increase gradually (*R^2=0.799*). This is completely in line with the trend in all education categories in the WoS database (Sezgin et al., 2022).

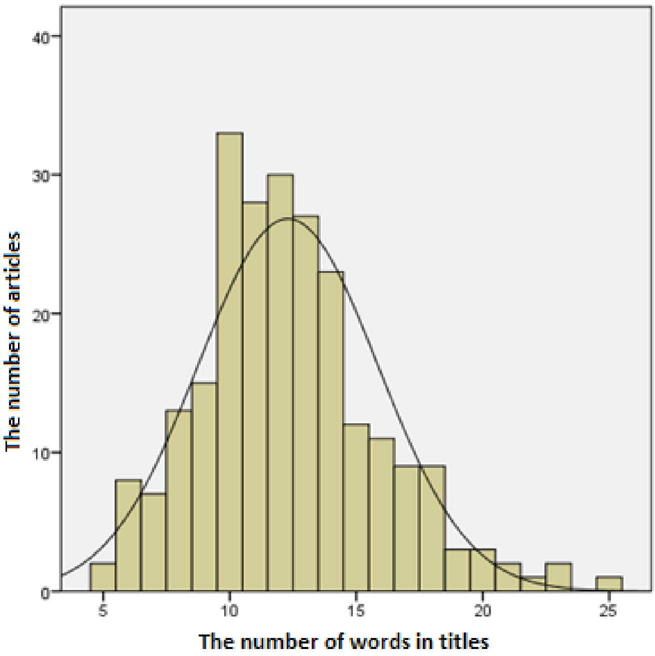
Figure 4. *The average of cited references in the issued articles through 2014-2021 annually.*



Many researchers often look at the title of an article to decide whether it is relevant to their studies or not. Hence, the first impression that the article title creates in the reader plays a major role whether the article will be read or not. Thus, the title is extremely important as it provides the most basic information about the content of the article (Hartley, 2008; Harzing, 2022). Letchford, Moat & Preis (2015) analyzed the 20.000 most cited articles published between 2007 and 2013, and found that articles with short titles received a higher number of citations. The reasons for this relationship are as follows: journals with high impact factors limit the word numbers in the title; new research on emerging topics has longer titles and is published in less prestigious journals due to the need for more explanation; short titles are easier to read and understand, and therefore, attract the reader more (Letchford, Moat & Preis, 2015).

Based on the obtained data, the article titles published in IJATE were analyzed. The mean value of the number of words in the titles was 12.33 (*median=12; sd=3.554; skewness=0.627; kurtosis=0.659*), and thus, it was identified that minimum 5 and maximum 25 words were used in the titles. It is understood that these values fit the normal distribution as seen in Figure 5. However, 10±3 is recommended for optimized article title length in the international literature (Elgendi, 2019).

Figure 5. *Title length distribution of articles published in IJATE.*



4. DISCUSSION and CONCLUSION

The outcomes reached by evaluating the publication performance of IJATE between 2014 and 2021 and the widespread impact of these publications are presented below.

The change in the number of authors, which is accepted as a measure of collaborative works in the field of education, is in consistence with the international literature. The international popularity of the journal has increased as listed in prestigious indexes and publishing the works of researchers from different countries. In the first years of the journal, its local appearance started to transform into a global one over time.

Recently, the use of online databases in academic publishing affects the widespread impact of articles, in other words, the number of citations they receive. This effect has made a significant positive contribution to the visibility of articles published in IJATE in the context of WoS.

It was concluded that the scope of the journal and the articles published were in complete harmony. However, it was found that the keywords of some articles were not selected in a desirable quality. On the other hand, it was concluded that the journals cited by IJATE or the journals citing by IJATE were varied.

It was found that the number of references used in the articles tended to increase in parallel with the international literature. However, it was concluded that the average title length of the articles was slightly longer than the ideal title length in the relevant literature.

Based on the discussion and conclusions, the following recommendations were provided.

- Researchers from the USA, Anglo-Saxon, and Continental European countries, which are prominent in the field of education both in terms of article productivity and the widespread impact of articles, should be encouraged to publish articles in the journal.
- Interactive applications can be developed for the visibility of articles on social academic networking sites such as Academia, ResearchGate and Linkedln, which will be established specifically for IJATE, to increase the positive significant relationship between the usage count of articles and the number of citations they receive.
- During the manuscript evaluation process, editors and/or referees should provide necessary guidance to authors by considering international trends for article titles and keywords.

Finally, it is thought that these suggestions will provide guidance to editors, reviewers and authors. Therefore, what has been done so far in IJATE should be a starting point to improve the current situation.

4.1. Limitations

There are a few limitations of this study, notwithstanding several crucial contributions. For one, bibliometric indicators based on the number of citations are time-dependent indicators and may change over time. Secondly, WoS and GS databases were used in the citation search and the study did not control for self-citations.

Acknowledgments

The authors would like to thank the blind reviewers for their useful comments and insightful suggestions.

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

Orhan Karamustafaoglu: Conceptualization, methodology, investigation, formal analysis, writing–original draft. **Metin Orbay:** Methodology, formal analysis, writing – review & editing. **Izzet Kara:** Formal analysis, review & editing.

Orcid

Orhan Karamustafaoglu https://orcid.org/0000-0002-2542-0998
Metin Orbay https://orcid.org/0000-0001-5405-2883
Izzet Kara https://orcid.org/0000-0002-9837-2819

REFERENCES

Aman, V., & Botte, A. (2017). A bibliometric view on the internationalization of European educational research. *European Educational Research Journal*, *16*(6), 843-868.

Aktaş, İ., & Karamustafaoğlu, O. (2022). Evaluation of the published articles in educational field: A Bibliometric analysis. *Hacettepe University Journal of Education, 37*(3), 1037-1050, https://doi.org/10.16986/HUJE.2021067584

Arslan, R., Orbay, K., & Orbay, M. (2022). Bibliometric Profile of an Emerging Journal: Participatory Educational Research. *Participatory Educational Research*, *9*(4), 153-171.

Budd, J.M., & Magnuson, L. (2010). Higher education literature revisited: Citation patterns examined. *Research in Higher Education, 51*(3), 294-304.

Chuang, K.Y., Huang, Y.L., & Ho, Y.S. (2007). A bibliometric and citation analysis of stroke-related research in Taiwan. *Scientometrics*, *72*(2), 201–212

Civera, A., Lehmann, E.E., Paleari, S., & Stockinger, S.A. (2020). Higher education policy: Why hope for quality when rewarding quantity? *Research Policy*, *49*(8), 104083.

Doğan, G., Dhyi, S.M.M.A., & Al, U. (2018). A Research on Turkey-addressed dropped journals from Web of Science. *Turkish Librarianship, 32*(3), 151-162.

Dong, B., Xu, G., Luo, X., Cai, Y., & Gao, W. (2012). A bibliometric analysis of solar power research from 1991 to 2010. *Scientometrics*, *93*(3), 1101-1117.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W.M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285-296.

Elgendi, M. (2019). Characteristics of a highly cited article: A machine learning perspective. *IEEE Access*, *7*, 87977-87986.

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's law in action. *GigaScience*, *8*(6), giz053.

Goodyear, R.K., Brewer, D.J., Gallagher, K.S., Tracey, T.J.G., Claiborn, C.D., Lichtenberg, J.W., & Wampold, B.E. (2009). The intellectual foundations of education: core journals and their impacts on scholarship and practice. *Educational Researcher, 38*(9), 700-706.

Hartley, J. (2008). *Academic writing and publishing: A practical handbook*. Routledge.

Harzing, A.W. (2007). *Publish or Perish,* available from https://harzing.com/resources/publish-or-perish Accessed November 30, 2022.

Harzing, A.W. (2022). *Publishing in academic journals: Crafting your career in academia*, Tarma Software Research Ltd, London, United Kingdom.

Henriksen, D. (2016). The rise in co-authorship in the social sciences (1980-2013). *Scientometrics, 107*(2), 455-476.

HIUC-Head of Inter-University Council (Üniversitelerarası Kurul Başkanlığı) (2022). The panel for the assessment of the position of Associate Professor in Fundamental Educational Sciences, https://www.uak.gov.tr/Documents/docentlik/2018-ekim-donemi/basvuru-sartlari/TA_Tablo1_2018E_071217.pdf (in Turkish). Accessed November 30, 2022.

Hicks, D. (2012). One size doesn't fit all: On the co-evolution of national evaluation systems and social science publishing. *Confero: Essays on Education Philosophy and Politics, 1*(1), 67–90.

Huang, D.W. (2016). Positive correlation between quality and quantity in academic journals. *Journal of Informetrics, 10*(2), 329-335.

IJATE (International Journal of Assessment Tools in Education) (2022). Available from https://dergipark.org.tr/en/pub/ijate Accessed November 30, 2022.

Letchford, A., Moat, H.S., & Preis T. (2015). The advantage of short paper titles. *Royal Society Open Science, 2*(150266), 1-6.

Liu, W., Hu, G., & Gu, M. (2016). The probability of publishing in first-quartile journals. *Scientometrics, 106*(3), 1273-1276.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & López-Cózar, E.D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160-1177.

Nemati-Anaraki, L., Zarghani, M., Ashoori-Mehranjani, F., & Eshaghi-Kopaei, S. (2019). Relationship between the usage count and the number of citations in the journals of Library and Information Sciences: The case of access type. *Library Philosophy and Practice (e-journal)*. 2376. https://digitalcommons.unl.edu/libphilprac/2376 Accessed November 30, 2022.

Orbay, K., Miranda, R., & Orbay, M. (2020). Building journal impact factor quartile into the assessment of academic performance: A case study. *Participatory Educational Research*, *7*(2), 1-13.

Orbay, M., Karamustafaoğlu, O., & Miranda, R. (2021). Analysis of the journal impact factor and related bibliometric indicators in education and educational research category. *Education for Information*, *37*(3), 315-336.

Özenç Uçak, N., & Al, U. (2008). Citation Characteristics of Social Sciences Theses. *Journal of Faculty of Letters*, *25*(2), 223–240.

Pajić, D., & Jevremov J. (2014) Globally national-locally international: Bibliometric analysis of a SEE psychology journal. *Psihologija, 47*(2), 263-267.

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world. *Publications*, *9*(1), 12. https://doi.org/10.3390/publications9010012

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation, 25*(4), 348-349.

Sezgin, A., Orbay, K., & Orbay, M. (2022). Educational research review from diverse perspectives: A Bibliometric analysis of Web of Science (2011-2020). *Sage Open*, *12*(4). https://doi.org/10.1177/21582440221141628

Tonta, Y. (2017). Journals published in Turkey and indexed in Web of Science: An evaluation. *Turkish Librarianship, 31*(4), 449-482.

ULAKBİM (Turkish Academic Network and Information Center) (2022). https://dergipark.org.tr/tr/pub/per , Accessed November 30, 2022.

Van Dalen, H.P. (2021). How the publish-or-perish principle divides a science: The case of economists. *Scientometrics*, *126*(2), 1675-1694.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping, *Scientometrics, 84*(2), 523-538.

Wang, X., Fang, Z., & Sun, X. (2016). Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count. *Scientometrics*, *109*(2), 917-926.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 12–28

https://doi.org/10.21449/ijate.1104005

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

A systematic literature review on multi-criteria decision making in higher education

Fatma Seyma Yuksel[1,*], Ayse Nilgun Kayadelen[2], Zahide Figen Antmen[3]

[1]Cukurova University, Faculty of Engineering, Department of Industrial Engineering, Adana, Türkiye
[2]Kutahya Dumlupinar University, Tavsanli Faculty of Applied Sciences, Department of Management Information Systems, Tavsanli, Kutahya, Türkiye
[3]Cukurova University, Faculty of Engineering, Department of Industrial Engineering, Adana, Türkiye

ARTICLE HISTORY

Received: Apr. 15, 2022

Accepted: Jan. 25, 2023

Keywords:

Multi-Criteria Decision-Making Techniques,

Higher education,

Literature review.

Abstract: The three components that form the basis of the educational process are the teacher, the learner, and the environment. These three components are affected by the developing and changing technology as a result of globalization considerably. Teaching and learning techniques should be updated and connected with these developments; new tools are therefore needed to make the necessary updates. Determination and application of the new tools include many decisions. Decision-makers can make more effective decisions using Multi-Criteria Decision-Making Techniques (MCDM), a complex decision-making tool that includes both quantitative and qualitative factors at present time. This study aimed to determine which MCDM methods are used in studies conducted in higher education, which is one of the most important development level indicators of countries, and to present a systematic literature review of MCDM method applications. The study was conducted in three stages: first, known electronics were searched until the end of 2021 using keywords; then, all studies were listed in a systematic taxonomy, and in the last stage, Thematic Network Analysis was used to evaluate the development of MCDM studies in the higher education area. It is determined that the Analytical Hierarchy Process (AHP) method is the most widely used method in higher education in MCDM applications. It was observed that the most common use of MCDM applications in higher education is e-learning as well. This study aims to be a guide for all researchers and practitioners who will study in both higher education and the MCDM areas.

1. INTRODUCTION

Higher education, also referred as post-secondary education or the third level, occurs after completing secondary education, is optional, and forms the final stage of formal education. Higher education includes many institutions and employees and a pretty high number of people benefit from such institutions as vocational schools, colleges, faculties, and institutions. The decision-making process in an institution becomes significant as the work in the institution gets intensive, the number of employees increases, and the institution's structure gets more complicated. The administration process, which begins with planning, emphasizes the necessity and importance of decision-making through research conducted eventually. The decision-

*Corresponding Author: Fatma Seyma Yuksel ✉ seymadnmz@gmail.com Cukurova University, Faculty of Engineering, Department of Industrial Engineering, Adana, Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

making process, vital to institutions, significantly affects administrators, employees, and people who demand a product and service from such institutions. Decision-makers may have to take more responsibilities as they make decisions on behalf of their institutions (Celikten et al., 2019). Therefore, decision-making is an essential process for higher education institutions.

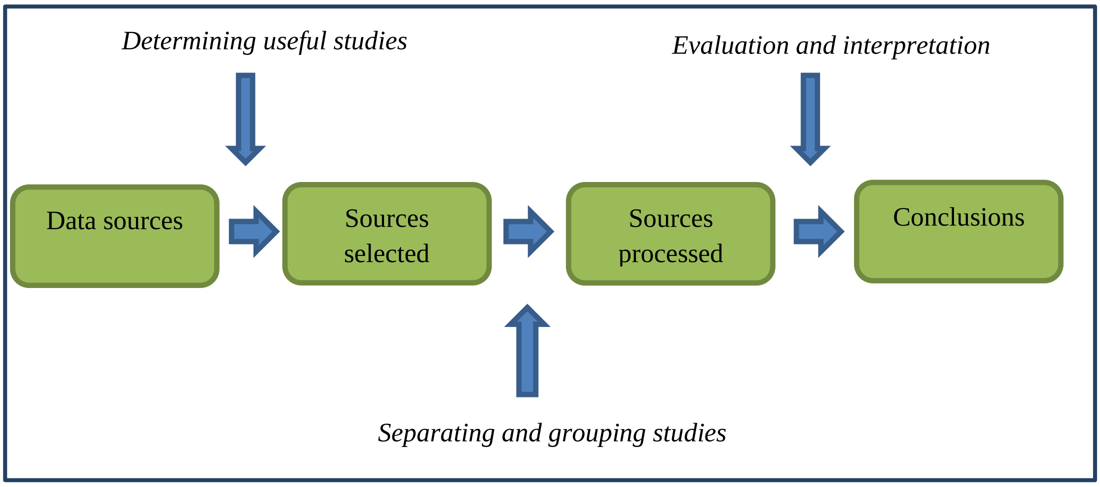
Academic and administrative staff selection, source selection, selection of learning systems, performance evaluation, etc. can be stated as examples of decision-making processes in higher education. Decisions to be made can be complex or straightforward, and their risk levels can be higher than expected. As the complexity and importance of a decision increase, the pressure on decision-makers and their importance increase as well. Here, decision-makers can select one or more alternatives among the others using various methods. Multi-Criteria Decision-Making (MCDM) techniques are the tools that provide decision-makers to make accurate, reliable, and quick decisions. For this reason, MCDM is among the most effective methods which decision-makers frequently apply and as a tool it ensures the best selection among alternatives adhering to multiple criteria used simultaneously (Mendoza & Prabhu, 2000). Since the 1960s, MCDM has been on the research agenda and many theories, many theoretical and applied articles, and books have been published in this field (Naveed et al., (2020); Rakesh et al., 2019; Roy, 2005). These techniques have an extensive application area, particularly in education, health, supply chain, transportation, computer science applications, energy, airway, banking, and production. MCDM techniques come to the forefront in case of a criterion and a goal that involves multiple qualities and quantities, particularly for educational institutions. Higher education institutions are rapidly and extensively affected by external factors such as social, economic, technological, or cultural (Timor, 2011).

This study aimed to review the literature that mainly focused on decision-making regarding the problems in higher education. The study is also thought to guide future studies to be conducted on MCDM problems in higher education.

Decision-making problems in higher education that were investigated by this study are commonly used for academic and administrative staff selection, graduate student selection, and e-learning systems selection. Studies on the Analytic Hierarchy Process (AHP) show that this technique is frequently used in the field of education (Chen et al., 2015; Sanchez et al., 2020; Yiğit et al., 2014). The studies examined also reveal AHP as the most frequently applied MCDM technique. In addition to AHP, the following techniques are used: Analytic Network Process (ANP), Fuzzy AHP (FAHP), VIse KriterijumsaOptimiz Ecija I Kompromisno Resenje (VİKOR), Technique for Order of Preference by Similarity to Ideal Solution (TOPSİS), Fuzzy TOPSİS (FTOPSIS), Complex Proportional Assessment (COPRAS), ELimination Et Choice Translating Reality (ELECTRE III), and also some hybrid techniques in which these techniques are used together (Anggrainingsih et al., (2018); Chen & Chen, 2010; Choi & Jeong, 2019; Giannoulis, 2010; Mazumdar, 2009; Naveed et al., 2020; Subbaiah et al., 2014).

In the in-depth analysis of the related literature, no study addressing MCDM techniques in higher education was encountered. To fill this gap in the literature, the present study, focusing on the MCDM and the latest related trends, reviewed publications on MCDM practices in higher education by the beginning of 2021. Figure 1 depicts the implementation stages of the present study.

Figure 1. *Study diagram.*


Determining useful studies
Evaluation and interpretation
Data sources
Sources selected
Sources processed
Conclusions
Separating and grouping studies

2. METHOD

In studies taken as reference, those using MCDM were primarily scanned considering the field of higher education. For this purpose, known databases (Web of Science, IEEE Xplore, Scopus, Science Direct, & Google Scholar) were searched using appropriate keywords by the beginning of 2021. "MCDM Techniques and Their Practices", "Higher Education", and "University" were examples of the keywords used. The search was done using different combinations of these keywords. Articles on MCDM practices in higher education were published by journals mainly in EBSCO, Google Scholar, ProQuest, SCImago, SCOPUS, Social Science Citation Index, Science Citation Index Expanded, Scopus, Emerging Sources Citation Index, Web of Science, and IEEE. Criteria for the articles to be included in this study were published either in English or in Turkish and in a journal or in a significant conference paper or in an article generated from a thesis.

Studies including MCDM practices outside of higher education, articles published in a language different from English (e.g., Arabic, French, etc.), and reviews that are not research articles were not included in the study. In line with these criteria, the present study was summarized in a table consisting of 72 scientific articles and conference papers (Table 1).

Table 1. *General literature table.*

Author Name	Year	MCDM Technique Used	Publication Type
Saaty & Ramanujam	1983	AHP	Journal
Liberatore &Nydick	1997	AHP	Journal
Kwak & Lee	1998	AHP	Journal
Drake	1998	AHP	Journal
Murakoshi et al.	2001	AHP	Conference
Aytaç & Bayram	2001	AHP	Journal
Özdemir & Gasimov	2004	AHP	Journal
Badri & Adulla	2004	AHP	Journal
Fenga et al.	2004	AHP	Journal
Grandzol	2005	AHP	Journal
Bali & Gencer	2005	AHP-FAHP	Journal
Kousalya et al.	2006	AHP	Journal
Begicevic & Divjak	2006	AHP & ANP	Journal
Colace et al.	2006	AHP	Conference
Ho et al.	2006	AHP	Journal
Tzeng et al.	2007	AHP & DEMATEL	Journal
Begicevic et al.	2007	AHP	Journal
Ray	2007	AHP	Journal

Table 1. *Continues*

Ho et al.	**2007**	**AHP**	**Journal**
Ozkul et al.	2007	AHP	Conference
Mustaffa et al.	2007	AHP	Journal
Tekindal & Erumit	2007	Classic Technique-AHP-FAHP	Journal
Begicevic et al.	2007	AHP	Journal
Melon et al.	2008	AHP & Direct Evaluation	Journal
Shee & Wang	2008	AHP	Journal
Chen & Chen	2008	VIKOR	Conference
Dundar	2008	AHP	Journal
Chi et al.	2008	FAHP	Journal
Nikoomaram et al.	2009	FAHP & FTOPSİS	Journal
Chao & Chen	2009	AHP (within CFPR)	Journal
Bo et al.	2009	FAHP	Conference
Ho et al.	2009	AHP	Journal
Lesmes et al.	2009	ANP	Conference
Sagir & Ozturk	2010	ANP	Journal
Altunok et al.	2010	Hybrid Model	Journal
Chen & Chen	2010	Hybrid Model	Journal
Gupta et al.	2010	AHP	Journal
Giannoulis & Ishizaka	2010	ELECTRE III	Journal
Lee	2010	AHP	Journal
Kara & Karaca	2010	AHP	Journal
Jie	2010	FAHP	Conference
Chen & Yang	2010	AHP	Conference
Lin	2010	FAHP	Journal
Mehregan et al.	2011a	FAHP	Journal
Mehregan et al.	2011b	FAHP	Conference
Nilashi & Janahmadi	2012	AHP	Journal
Li et al.	2012	AHP	Conference
Soba	2012	AHP	Journal
Wu et al.	2012	Hybrid Model	Journal
Syamsuddin	2012	FAHP	Journal
Kiarazm & Koohkan	2013	AHP	Journal
Kurilovasa & Zilinskiene	2013	Hybrid Model	Journal
Ozturk	2014	ANP	Journal
Yigit et al.	2014	AHP	Journal
Subbaiah et al.	2014	TOPSİS	Journal
Omurbek et al.	2014	TOPSİS+VİKOR based on AHP	Journal
Aly et al.	2014	AHP-TOPSİS	Journal
Mondal & Pramanik	2014	Hybrid Model	Journal
Nagpal et al.	2015	FAHP	Journal
Chen et al.	2015	AHP	Journal
Jain et al.	2016	AHP-TOPSİS	Journal
Garg	2017	FAHP, WEDBA, COPRAS	Journal
Garg & Jain	2017	FAHP, COPRAS, VIKOR, WDBA	Journal
Naveed et al.	2017	FAHP	Conference
Kabak et al.	2017	Hybrid Model	Journal
Ghosh & Pal	2017	Hybrid Model	Journal
Cebi & Karal	2017	FAHP	Journal
Anggrainingsih et al.	2018	FAHP	Conference
Mohammed et al.	2018	FAHP-TOPSİS	Journal
Choi & Jeong	2019	ANP	Journal
Garg et al.	2019	COPRAS-F	Conference
Naveed et al.	2020	AHP & FAHP	Journal

3. RESULT

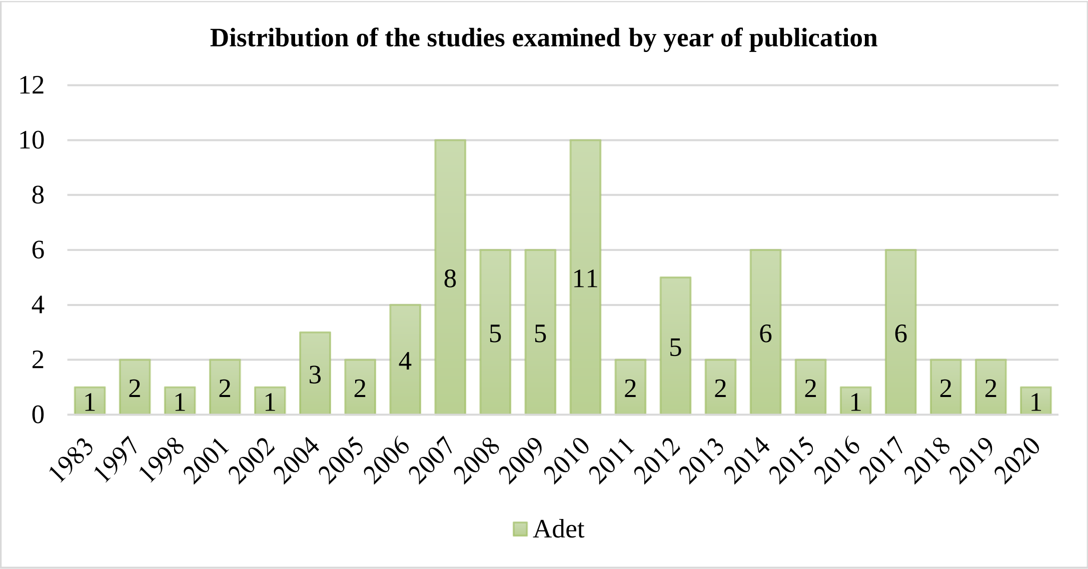
3.1. Classification of MCDM Techniques in Higher Education

In this section, the reference articles were grouped according to these characteristics: (1) Publication year, (2) MCDM technique applied, (3) Application Field, (4) Publication Type, and (5) Index Scanned.

3.1.1. *Publication year*

Among the reference articles according to their publication year, the first study on MCDM practice in higher education was carried out in 1983. The most up-to-date study in the application field was carried out in 2020. Besides, an increase in the number of articles published was observed between 2007–2010 and 2014–2017 (Figure 2).

Figure 2. Distribution of the reference studies by year of publication.



3.1.2. *MCDM technique applied*

Studies on MCDM practices in higher education used various MCDM techniques. In the reference studies, AHP, FAHP, ANP, TOPSIS, FTOPSIS, VIKOR, etc. techniques and the hybrid techniques combining these techniques were used. As shown in the graph of the distribution of the methods used in Figure 3, AHP is the most frequently used one among all the techniques (Table 1). The AHP technique, developed by Thomas Saaty in 1980, is an effective MCDM tool to help decision-makers determine the best choice while making complicated decisions. An essential characteristic of the AHP technique is comparing alternatives based on various criteria and using pairwise comparisons to predict criterion weights. AHP is based on the priority theory. At the core of AHP, a systematic approach is followed for an alternative selection and justification problem using the fuzzy set theory and hierarchical structure analysis. Since it has a fuzzy basis, this technique can be used in situations where user preference is identified (Aruldoss et al., 2013). In the first MCDM practice in Higher Education (Saaty & Ramanujam, 1983), an evaluation was made regarding staff selection in universities using AHP. With this study, 35 reference studies used AHP for MCDM practices in higher education. In this field, Chen et al. (2010) carried out the most up-to-date study in 2015 on determining the educational quality level of administrators (Table 2).

Figure 3. *Distribution of the reference studies by the technique applied.*

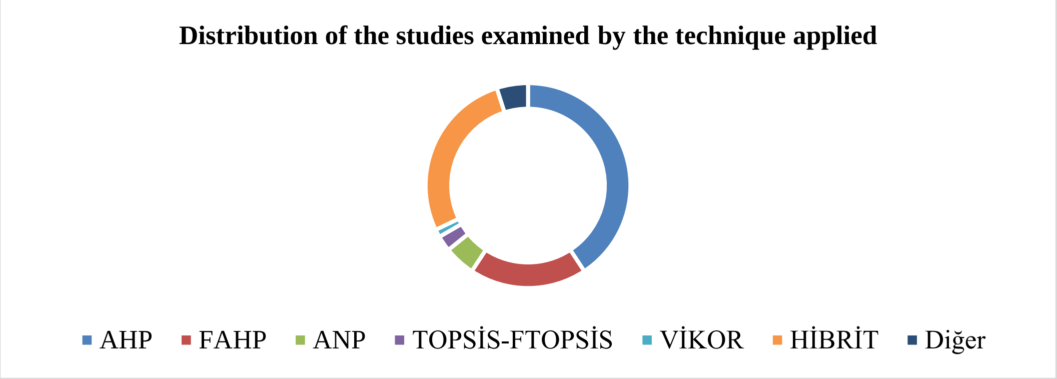


Table 2. *Studies applying AHP*.

Author Name	Year	Application Field	Publication Type
Saaty & Ramanujam	1983	Staff Selection	Journal
Liberatore & Nydick	1997	Academic Research Articles	Journal
Kwak & Lee	1998	Source Distribution	Journal
Drake	1998	Engineering Major Selection	Journal
Murakoshi et al.	2001	e-Learning Comparison in formal education	Conference
Aytac & Bayram	2001	Undergraduate Student	Journal
Ozdemir & Gasimov	2004	Faculty Course Assignment System	Journal
Badri & Adulla	2004	Higher Education Performance	Journal
Fenga et al.	2004	Performance Evaluation in Universities	Journal
Grandzol	2005	Staff Selection	Journal
Kousalya et al.	2006	Student Absence	Journal
Colace et al.	2006	e-Learning Platforms	Conference
Begicevic et al.	2007	e-Learning	Journal
Ray S.	2007	Thesis Advisor Selection	Journal
Ho et al.	2007	Source Distribution	Journal
Ozkul et al.	2007	Distance Education	Conference
Mustaffa et al.	2007	Academic Staff	Journal
Begicevic et al.	2007	e-Learning	Journal
Melón et al.	2008	Face-to-Face and Online Education	Journal
Shee & Wang	2008	Student Satisfaction	Journal
Dundar	2008	Course Selection	Journal
Chao & Chen	2009	e-Learning	Journal
Ho et al.	2009	e-Learning	Journal
Lee	2010	Performance Evaluation in Universities	Journal
Kara & Karaca	2010	Determining criteria that are effective in department selection	Journal
Chen & Yang	2010	e-Learning	Conference
Gupta et al.	2010	Education Evaluation	Journal
Nilashi & Janahmadi	2012	e-Learning	Journal
Li et al.	2012	e-Learning	Conference
Soba	2012	Higher Education	Journal
Kiarazm & Koohkan	2013	Performance Evaluation	Journal
Yigit et al.	2014	Course Content	Journal
Chen et al.	2015	University Administrators	Journal

Van Laarhoven and Pedrycz carried out the first theoretical study on FAHP in 1983. FAHP emerged due to the combination of the fuzzy relationship and pairwise comparison concepts (Toksarı, 2011). Unlike AHP, which uses clear values, comparison ratios are given within a range in FAHP (Şengül et al., 2013). FAHP is an effective tool for decision-making processes that cannot be quantified with specific data and where uncertainty is great. In this approach, decision-makers are asked to verbally express their evaluation at the stage of identifying the criteria weights. With this aspect, FAHP is a more realistic evaluation technique (Kusakci, 2019). 11 reference studies used the FAHP technique: the first was carried out by Chi et al. (2008) to evaluate higher education departments in the Ministry of Education in Taiwan; the most up-to-date of those studies was a conference paper by Anggrainingsih et al. (2018) using an e-learning FAHP application (Table 3).

Table 3. *Literature table for the studies applying AHP.*

Author Name	Year	Application Field	Publication Type
Chi et al.	2008	Development of University Organizations	Journal
Bo et al.	2009	e-Learning	Conference
Lin	2010	e-Learning	Journal
Jie	2010	e-Learning	Conference
Mehregan et al.	2011a	Evaluating e-learning performance	Journal
Mehregan et al.	2011b	e-Learning	Conference
Syamsuddin	2012	e-Learning Software	Journal
Nagpal et al.	2015	Evaluation of University Websites	Journal
Naveed et al.	2017	e-Learning	Conference
Cebi & Karal	2017	Student Projects	Journal
Anggrainingsih et al.	2018	e-Learning	Conference

ANP is a convenient MCDM technique coined and developed by Thomas L. Saaty (1999) and used to calculate weights and priorities. ANP is a general form of AHP used to consider non-hierarchical structures in MCDM (Chen, 2010). This technique demonstrates problems as networks by determining the relationship between their components (Omurbek, 2014). ANP considers the between- and in-group dependencies and feedback between the criteria. With this characteristic, ANP facilitates the solution of decision-making problems more effectively and realistically (Bo et al., 2009; Goksu, 2008; Jie, 2010; Lin, 2010). Most of the ANP studies were applied to e-learning. Table 4 shows the reference studies using the ANP technique. The first study using ANP was a university program application by Lesmes et al. in 2009. The most up-to-date study was an e-learning application carried out by Choi and Jeong in 2019.

Table 4. *Literature table for the studies applying AHP.*

Author Name	Year	Application Field	Publication Type
Lesmes et al.	2009	University Program	Conference
Sagir & Ozturk	2010	Observer, Exam	Journal
Ozturk	2014	Open and Distance Education System	Journal
Choi & Jeong	2019	e-Learning	Journal

TOPSIS, another method used in this field, is an MCDM technique developed by Hwang and Yoon (1981). Evaluation of alternatives (decision choices) is based on two main points; namely, positive ideal solution and negative ideal solution. In the TOPSİS technique, the target is the decision choice closest to the positive ideal solution and farthest to the negative ideal solution. The positive ideal solution is the solution that makes the cost criterion minimum and the benefit criterion maximum. Comparatively, the negative solution is the solution that makes the cost criterion maximum and the benefit criterion minimum. There must exist at least two decision choices to apply this technique (Altunok, 2010). The TOPSİS technique is quite simple and understandable, and effective in calculations. It determines the relative performance of alternatives with simple mathematical formulas (Kabak, 2017). In our review study, one reference study used the TOPSİS technique as Subbaiah et al. (2014) determined the criteria for the ranking and evaluation of engineering and education institutes using TOPSİS (Table 5).

Table 5. *Literature table for the studies applying TOPSİS.*

Author Name	Year	Application Field	Publication Type
Subbaiah et al.	2014	Ranking and Evaluation of Engineering and Education Institutes	Journal

VIKOR, another MCDM technique, is a decision-making technique suggested by Opricovic and Tzeng (2004) to solve multi-criteria problems in complex systems; namely, the systems consisting of criteria that might contradict each other (Tezergil, 2016). VIKOR is based on the combination function representing the solution closest to the ideal solution (Opricovic, 1998). The VIKOR technique is mainly used in situations where decision-makers cannot make a selection determinedly or explain their choice (Paksoy, 2015) (Table 6). Our review of the related literature shows that only Chen and Chen (2008) used the VIKOR technique to determine the selection criteria of university type.

Table 6. *Literature table for the studies applying VIKOR.*

Author Name	Year	Application Field	Publication Type
Chen & Chen	2008	University Type Selection	Conference

ELECTRE (Elimination Et Choix Traduisant la Realité), another MCDM technique, is math-based and used for optimization. The ELECTRE III technique developed by Bernard Roy in 1978 is used in ranking problems. The technique selects the best choice among the alternatives asked to be evaluated and ranks the remaining options from the most optimal to the least optimal (Keles, 2019) (Table 7). The reference article by Giannoulis and Ishizaka (2010) used the ELECTRE III technique to compare English Universities.

Table 7. *Literature table for the studies applying ELECTRE III.*

Author Name	Year	Application Field	Publication Type
Giannoulis & Ishizaka	2010	Comparison of English Universities	Journal

COPRAS-F, another MCDM technique, has been obtained by combining fuzzy logic and the classic COPRAS technique to cope with the inability to make effective decisions due to ambiguities. In the COPRAS-F technique, performance values consisting of fuzzy numbers are used. The technique benefits from linguistic scales (Çakir & Ozdemir, 2018). The technique first suggested by Zavadskas and Kaklauskas in 1996 is based on selecting the best alternative by determining a ratio with the positive and negative optimal solution (Yazdani et al., 2011). There was one reference conference paper that used the COPRAS-F technique; namely, Garg et al. (2019) used COPRAS-F in the field of e-learning (Table 8).

Table 8. *Literature table for the studies applying COPRAS-F.*

Author Name	Year	Application Field	Publication Type
Garg et al.	2019	e-Learning	Conference

Some reference studies also used the hybrid technique that combined multiple MCDM techniques. Hybrid studies used the following combination of MCDM techniques: AHP-ANP, AHP-DEMATEL, FAHP-FTOSİS, AHP-Quality Function Deployment (QFD), AHP-Weighted Product (WP) -TOPSİS, The Decision Making Trial And Evaluation Laboratory (DEMATEL)-FANP-TOPSİS, AHP-VİKOR, Multiple Criteria Evaluation of the Quality of Learning Software (MCEQLS)-AHP, TOPSİS-VİKOR, AHP-TOPSİS, Multicriteria Group Decision Making (MCGDM), FAHP- WeighBalited Euclidean Distance Based Approach (WEDBA)-COPRAS, FAHP-COPRAS-VIKOR- Weighted Distance Based Approximation (WDBA), and ANP-TOPSİS. Of the reference studies, 19 used hybrid techniques. Bali O. and Gencer conducted the first study with hybrid techniques in 2005. In their study, they used AHP and FAHP together. Naveed Q. N. et al. made the most up-to-date hybrid-technique study on e-learning in 2020 (Table 9).

Table 9. *Literature table for the studies applying hybrid techniques.*

Author Name	Year	Application Field	Publication Type
Bali & Gencer	2005	Student Selection	Journal
Begicevic & Divjak	2006	e-Learning	Journal
Tzeng et al.	2007	e-Learning	Journal
Tekindal & Erumit	2007	Graduate Student Selection	Journal
Nikoomaram et al.	2009	Performance Evaluation	Journal
Altunok et al	2010	Graduate Student	Journal
Chen & Chen	2010	Innovation Support System	Journal
Wu et al.	2012	Performance Evaluation in Universities	Journal
Kurilovasa & Zilinskiene	2013	e-Learning Quality Assessment	Journal
Omurbek et al.	2014	Performance Evaluation in Universities	Journal
Aly et al.	2014	Prioritizing Performance Indicators in Engineering Education	Journal
Mondal & Pramanik	2014	Staff Selection	Journal
Jain et al.	2016	e-Learning	Journal
Garg	2017	e-Learning	Journal
Garg & Jaina	2017	e-Learning	Journal
Kabak et al.	2017	University	Journal
Ghosh & Pal	2017	Academic Performance	Journal
Mohammed et al.	2018	e-Learning	Journal
Naveed et al	2020	e-Learning	Journal

Application Field: MCDM techniques offer essential decision-making tools for the process in higher education. Considering the reference studies in terms of the application field, e-learning ranks first. Staff selection, performance evaluation, source distribution, and course selection were among the other application fields.

Publication Type: The reference studies in this study are 72 in total, 59 of which are articles and 13 of which are scientific conference proceedings.

Index scanned: Of the reference articles, 59 were published in journals reviewed by significant indices like EBSCO, Google Scholar, ProQuest, SCImago, SCOPUS, Social Science Citation Index (SSCI), Science Citation Index Expanded (SCIE), Science Citation Index (SCI), Emerging Sources Citation Index (ESCI), Web of Science, and Institute of Electrical and

Electronics Engineers (IEEE). Five reference articles were published in journals of SSCI, seven in journals of SCIE, two in journals of SCI, and six in journals of ESCI. Furthermore, seven reference articles were published in journals of IEEE indices. Other reference studies were published in common indices like EBSCO, Google Scholar, ProQuest, SCImago, SCOPUS, and Web of Science.

The present study made a literature review of all studies on MCDM practices in higher education until today. As a result of this literature review, no up-to-date research on this field was encountered. In this regard, the present study is thought to pioneer this field. By putting forward application fields in terms of decision-making in higher education, the present study is believed to direct future studies.

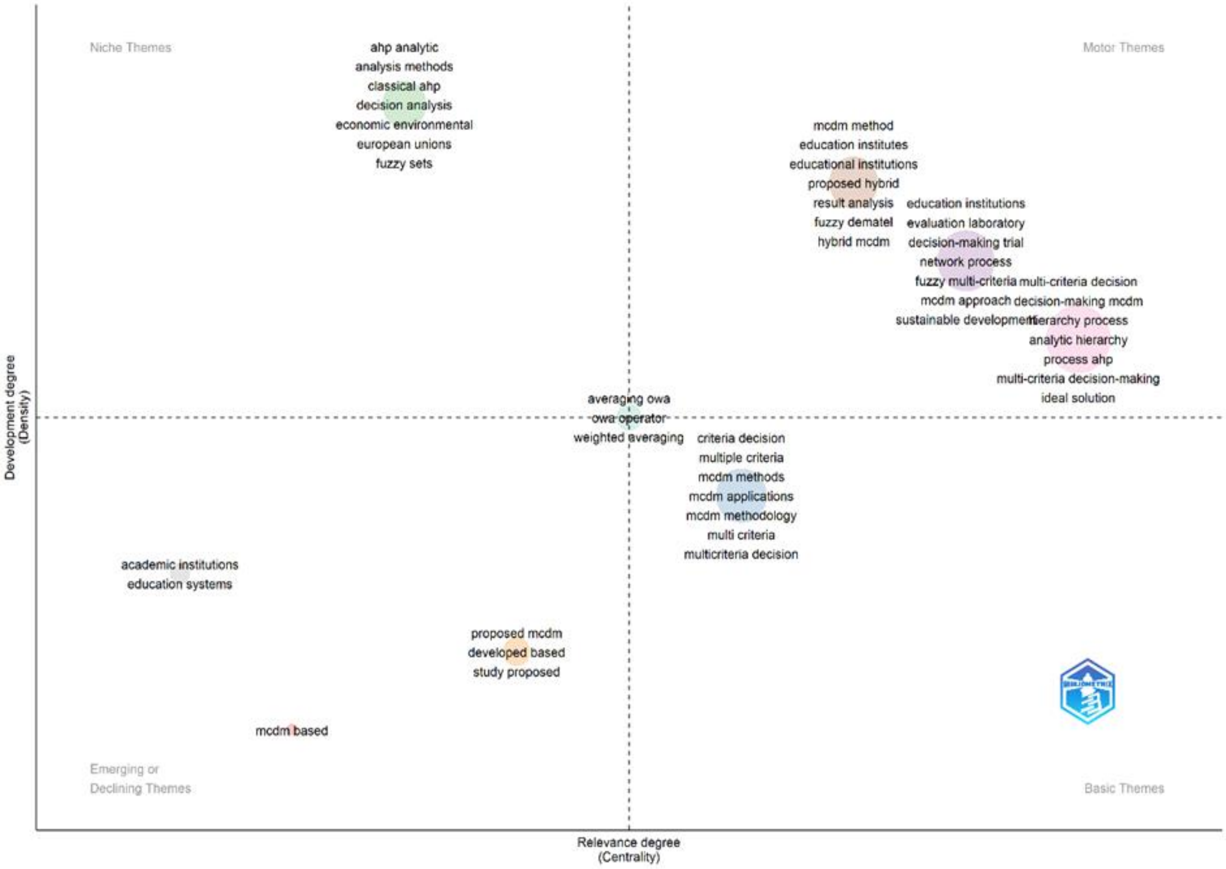
3.2. Thematic Network Analysis

Applying thematic networks is simply a way of organizing a thematic analysis of qualitative data. Thematic Analysis as a method was first developed by Gerald Holton, a physicist, and historian of science in the 1970s (Holton, 1975). The thematic analysis seeks to unearth the themes salient in a text at different levels, and thematic networks aim to facilitate the structuring and depiction of these themes. Thematic Analysis is a method for systematically identifying, organizing, and offering insights into patterns and meanings (themes) across a dataset.

In this study, Thematic Network Analysis was applied to evaluate the development of MCDM studies in the higher education area. The studies assessed by the R-based Bibliometrix Software were mapped thematically (Aria & Cuccurullo, 2017). Four categories; namely, engine, specialty, emerging, and basic themes used to categorize graphs in Thematic Mapping Engine themes are a group of themes that have strong links to other well-developed sub-themes. The most frequently used themes are in the engine themes. Niche themes include well-developed themes that are important in the research field. Emerging themes are low-density less advanced themes. Basic themes include fundamental ideas.

The thematic network map of studies in the higher education area is given in Figure 4. When Figure 4 is examined, it can be said that institutes of higher education are among the most studied themes. Thematic mapping, of which AHP is among the most widely used methods, is also emerging. In addition, it is observed that Fuzzy MCDM models are among the frequently used themes. MCDM studies related to education systems are not among the popular themes. Determination of criteria and evaluation of criteria are among the major themes.

Figure 4. *Thematic network of multi-criteria decision-making in higher education.*



4. DISCUSSION and CONCLUSION

Making accurate decisions is important for higher education institutions that include many intuitions and many employees. MCDM techniques are the tools that help decision-makers make accurate, reliable, and quick decisions. Therefore, it is suggested that MCDM techniques be used in the process of making quick decisions.

The present study addressed the decision-making process in higher education through the lens of MCDM. For this purpose, it presented 72 reference studies that discussed the trends in MCDM techniques in higher education. All the articles were classified by (1) publication year, (2) MCDM technique applied, (3) application field, (4) publication type, and (5) index scanned. Of the 72 articles examined, 35 used the AHP technique. Of the remaining studies, 19 used hybrid techniques and 11 used FAHP. One study for each of the TOPSİS, VİKOR, ELECTRE III, and COPRAS-F techniques was implemented in the higher education area. The AHP technique was the most common technique among MCDM practices in higher education, with 30 scientific articles and five conference papers. The AHP technique was the most common one among MCDM practices in higher education, with 30 scientific articles and 5 conference papers. Thematic Network Analysis also confirms all these results exposed in the study.

Overall, 11 studies, six scientific articles, and five conference papers applied MCDM in e-learning. Thus, e-learning was the most common application field. Due to the spread of the COVID-19 pandemic, universities in many different parts of the world have paused face-to-face education and continued their educational activities online. Therefore, e-learning has become very important. At such a time, MCDM techniques are foreseen to become essential tools to improve e-learning. The articles included in this study were scientific and published in journals that are scanned by significant indices. Of the articles made on MCDM practices in higher education, 15 were conducted in Türkiye. There existed seven articles that used the AHP method in the field of higher education in Türkiye.

The present study aimed to gather studies on MCDM practices in higher education and guide future researchers in this field. Therefore, future studies can be conducted using MCDM techniques different from those used by the existing studies. Concerning the importance of decision-making today, the authors of the present study hope for an increase in the number of MCDM techniques and approaches in the related literature.

Declaration of Conflicting Interests and Ethics

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

Authorship Contribution Statement

Fatma Seyma Yuksel: Design, Data Collection and/or Processing, Analysis and/or Interpretation and Writing. **Ayse Nilgun Kayadelen:** Concept, Data Collection and/or Processing, Analysis and/or Interpretation and Writing. **Zahide Figen Antmen:** Data Collection and/or Processing and Critical Review. All authors read and approved the final manuscript.

Orcid

Fatma Seyma Yuksel https://orcid.org/0000-0002-8080-2665
Ayse Nilgun Kayadelen https://orcid.org/0000-0002-5442-893X
Zahide Figen Antmen https://orcid.org/0000-0001-8475-1300

REFERENCES

Altunok, T., Özpeynirci, O., Kazancoglu, Y., & Yilmaz, R. (2010). Comparatives of multicriteria decisions making methods for postgraduate student selection. *Egitim Arastirmalari-Eurasian Journal of Educational Research*, *40*, 1-15.

Aly, M.F., Attia, H.A., & Mohammed, A.M. (2014). Prioritizing faculty of engineering education performance by using AHP-TOPSİS and balanced scorecard approach. *International Journal of Engineering Science and Innovative Technology*, *3*(1), 11-23.

Anggrainingsih, R., Umam, M.Z., & Setiadi, H. (2018). Determining e-learning success factor in higher education based on user perspective using Fuzzy AHP. *MATEC Web Conferences*. 154, 03011. https://doi.org/10.1051/matecconf/201815403011

Aytaç, S., & Bayram, N. (2001). Üniversite gençliğinin iş ve eş seçimindeki etkin kriterlerinin analitik hiyerarşi süreci (AHP) ile analizi [Analysis of university youth's effective criteria for job and spouse selection by analytical hierarchy process (AHP)]. *Öneri Dergisi*, *4*(16), 89-100. https://doi.org/10.14783/maruoneri.727643

Badri, M.A., & Abdulla, M.H. (2004). Awards of excellence in institutions of higher education: an AHP approach. *International Journal of Educational Management*, *18*(4), 224-242. https://doi.org/10.1108/09513540410538813

Bali, O., & Gencer, C. (2005). AHP Bulanık AHP ve Bulanık Mantıkla Kara Harp Okuluna öğretim elemanı seçimi [Ahp, Fuzzy Ahp, and Fuzzy Logic Selection of Academic Staff to Turkish Military Academy]. *Kara Harp Okulu Savunma Bilimleri Dergisi*, *4*, 24-43.

Begicevic, N., & Divjak, B. (2006). Validation of theoretical model for decision making about e-learning implementation. *Journal of Information and Organizational Sciences*, *30*(2), 171-184.

Begicevic, N., Divjak, B., & Hunjak, T. (2007). Development of AHP based-model for decision making on e-learning implementation. *Journal of Information and Organizational Sciences*, *31*, 13-24.

Bo, L., Xuning P., & Bingquan B. (2009). Modeling of network education effectiveness evaluation in fuzzy analytic hierarchy process. *International Conference on Networking and Digital Society, 2*, 198–200. ICNDS'09, IEEE.

Cakir, E., & Ozdemir, M. (2018). Altı sigma projelerinin bulanık copras yöntemiyle değerlendirilmesi: Bir üretim işletmesi örneği [Evaluation of six sigma projects with fuzzy copras method: An example of a manufacturing company]. *Verimlilik Dergisi, 1,* 7-39.

Cebi, A., & Karal, H. (2017). An application of fuzzy analytic hierarchy process (FAHP) for evaluating students' Projects. *Educational Research and Reviews*, *12*(3), 120-132. https://doi.org/10.5897/ERR2016.3065

Celikten, M., Gilic, F., Celikten., & Yildirim, A. (2019). Örgüt yönetiminde karar verme süreci: Bitmeyen bir tartışma [Decision making process in organization management: An endless discussion]. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 15*(2), 581-592.

Chao, R.J., & Chen, Y.H. (2009). Evaluation of the criteria and effectiveness of distance e-learning with consistent fuzzy preference relations. *Expert Systems with Applications, 36*, 10657-10662. https://doi.org/10.1016/j.eswa.2009.02.047

Chen, J.F., Hsieh, H.N., & Do, Q, H. (2015). Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach. *Applied Soft Computing, 28,* 100-108. https://doi.org/10.1016/j.asoc.2014.11.050

Chen, J.K., & Chen, I.S. (2010). Using a novel conjunctive MCDM approach based on DEMATEL, fuzzy ANP, and TOPSIS as an innovation support system for Taiwanese higher education. *Expert Systems with Applications, 37,* 1981-1990. https://doi:10.1016/j.eswa.2009.06.079

Chen, Y., & Yang, M. (2010). Study and construct an online self-learning evaluation system model based on the AHP method. *2nd IEEE International Conference on Information and Financial Engineering (ICIFE)*, 54–58.

Chena, J.F. Hsieha, H.N. & Do, Q.H. (2015). Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach. *Applied Soft Computing*, *28*, 100-108. https://doi.org/10.1016/j.asoc.2014.11.050

Chi, H.K., Yeh, H.R., & Liao, L.-H., (2008). Applying fuzzy analytic hierarchy process to explore the university organizational performance in Taiwan. *The Journal of Human Resource and Adult Learning*, *4*(1), 39–46.

Choi, C.R., & Jeong, H.Y. (2019). Quality evaluation for multimedia contents of e-learning systems using the ANP approach on a high-speed network. *Multimedia Tools and Applications*, *78*, 28853-28875. https://doi.org/10.1007/s11042-019-7351-8

Cicekli, U.G., & Karacizmeli, A. (2013). Bulanık analitik hiyerarşi süreci ile basarılı öğrenci seçimi: Ege Universitesi İktisadi ve İdari Bilimler Fakültesi örneği [Successful student selection with fuzzy analytic hierarchy process: The example of Ege University Faculty of Economics and Administrative Sciences]. *Ege Strategic Research Journal*, *4*(1), 77-103. https://doi.org/10.18354/esam.81730

Colace, F., Santo, M.D., & Pietrosanto, A. (2006). Evaluation models for e-learning platform: An AHP approach. *Proceedings of Frontiers in Education.* 1-6. https://doi.org/10.1109/FIE.2006.322312

Drake, P.R. (1998). Using the analytic hierarchy process in engineering education. *Int. J. Engng Ed.*, *14*, 191-196.

Dundar, S. (2008). Ders seçiminde analitik hiyerarşi prosesi uygulaması [Analytical Hierarchy Process application in course selection]. *Suleyman Demirel Universitesi Iktisadi ve Idari Bilimler Fakultesi Dergisi 13*(2), 217-226.

Ertugrul, I., & Karakasoglu, N. (2007). Fuzzy TOPSIS method for academic member selection in engineering faculty. *Innovations in E-learning, Instruction Technology, Assesment, and Engineering Education*, 151-156.

Feng, Y.J., Lu, H., & Bi, K. (2004). An AHP/DEA method for measurement of the efficiency of R&D management activities in universities. *Intl. Trans. In Op. Res.*, *11*, 181-191. https://doi.org/10.1111/j.1475-3995.2004.00450.x

Garg, R., & Jain, D. (2017). Fuzzy multi-attribute decision-making evaluation of e-learning websites using FAHP, COPRAS, VIKOR, WDBA. *Decision Science Letters*, *6*, 351-364. http://dx.doi.org/10.5267/j.dsl.2017.2.003

Garg, R. (2017). Optimal selection of E-learning websites using multiattribute decision-making approaches. *J. Multi-Criteria Decision Analysis*, *24*, 187-196. https://doi.org/10.1002/mcda.1612

Garg, R., Kumar, R., & Garg, S. (2019). MADM-Based parametric selection and ranking of e-learning websites using fuzzy COPRAS. *IEEE Trans. Educ.*, *62*(1), 11–18. https://doi.org/10.1109/TE.2018.2814611

Ghosh, D., & Pal, A. (2017). Analysis of faculty teaching using a multi-criteria decision-making approach. *International Journal of Engineering & Technology*, *7*, 74-78. https://doi.org/10.14419/ijet.v7i2.28.12884

Giannoulis, C., & Ishizaka, A. (2010). A Web-based decision support system with ELECTRE III for a personalized ranking of British universities. *Decision Support Systems*, *48*, 488-497. https://doi.org/10.1016/j.dss.2009.06.008

Grandzol, J.R. (2005). Improving the faculty selection process in higher education: A case for the analytic hierarchy process. *IR Applications*, *6*, 13.

Gupta, R., Garg, T.K., Gupta, S., & Goel, A. (2010). Decision Analysis Approach for Quality in Technical Education. *Global Journal of Human Social Science*, *10*(1), 14-18.

Ho, W., Dey, P.K., & Higson, H.E. (2006). Multiple criteria decision-making techniques in higher education. *International Journal of Educational Management*, *20*(5), 319-337. https://doi.org/10.1108/09513540610676403

Holton, G. (1975). On the role of themata in scientific thought. *Science*, 188(4186), 328-334.

Hsu, C.M., Yeh, Y.C., & Yen, J. (2009). Development of design criteria and evaluation scale for web-based learning platforms. *International Journal of Industrial Ergonomics*, *39*, 90-95. https://doi.org/10.1016/j.ergon.2008.08.006

Jain, D., Garg, R., & Bansal, A. (2016). Selection and ranking of E-learning websites using weighted distance-based approximation. *Journal of Computer Education*, *3*(2), 193-207. https://doi.org/10.1007/s40692-016-0061-6

Jesus. E.N., Rodrigues, J.C., & Antunes, C.H. (2007). A multicriteria decision support system for housing evaluation. *Decision Support Systems*, *43*, 779-790. https://doi.org/10.1016/j.dss.2006.03.014

Jie, C. (2010). Evaluation and modeling of online courses using fuzzy AHP. *2010 International Conference on Computer and Information Application*, 232-235.

Kabak, M., Ozceylan, E., Dagdeviren, M., & Genc T. (2017). Evaluation of distance education websites: a hybrid multicriteria approach. *Turkish Journal of Electrical Engineering & Computer Sciences*, *25*, 2809–2819. https://doi.org/10.3906/elk-1512-271

Kara, M., & Karaca, Y. (2010). Üniversite öğrencilerin işletme bölümünü seçmelerinde etkili olan öncelikli faktörlerin analitik hiyerarşi prosesi metodu ile analizi: Bozok Üniversitesi İktisadi ve İdari Bilimler Fakültesinde bir uygulama [Analysis of the priority factors that affect university students' choice of business administration with the analytical hierarchy process method: An application in Bozok University Faculty of Economics and Administrative Sciences]. *Organizasyon ve Yonetim Bilimleri Dergisi*, *2*(1), 133-140.

Kiarazm, A., & Koohkan, F. (2013). Performance evaluation in higher education institutes with the use of combinative model AHP and BSC. *Journal of Basic and Applied Scientific Research*, *3*(4), 940-944

Kousalya, P., Ravindranath, V., & Vizayakumar, K. (2006). Student absenteeism in engineering colleges: Evaluation of alternatives using AHP. *Journal of Applied Mathematics and Decision Sciences* 2006. 1–26. https://doi.org/10.1155/JAMDS/2006/58232

Kurilovas, E., & Zilinskiene, I. (2013). New MCEQLS AHP method for evaluating the quality of learning scenarios. *Technological and Economic Development of Economy*, *19*(1), 78-92. https://doi.org/10.3846/20294913.2012.762952

Kurilovas, E., & Serikoviene, S. (2013). New MCEQLS TFN method for evaluating quality and reusability of learning objects. *Technological and Economic Development of Economy*, *19*(4), 706-723. https://doi.org/10.3846/20294913.2013.837112

Kwak N.K., & Lee C. (1998). A multicriteria decision-making approach to university resource allocations and information infrastructure planning. *European Journal of Operational Research*, *110*(2), 234-242. https://doi.org/10.1016/S0377-2217 (97)00262-2

Lee, S.H. (2010). Using fuzzy AHP to develop intellectual capital evaluation model for assessing their performance contribution in a university. *Expert Systems with Applications*, *37*, 4941-4947. https://doi.org/10.1016/j.eswa.2009.12.020

Lesmes, D., Castillo, M., & Zarama, R. (2009). Application of The Analytic Network Process (ANP) to Establish Weights in Order to Re-Accredit a Program of a University. *Proceedings of the International Symposium on the Analytic Hierarchy Process*, 29.

Li, W., Gao, X., & Fu, G. (2012). Fuzzy comprehensive assessment of network environment and learning quality combined with the analytic hierarchy process. *2nd International Conference on Consumer Electrics, Communications and Networks*, 2600-2603, *IEEE*.

Liberatore, M.J., & Nydick, R.L. (1997). Group Decision Making in Higher Education Using the Analytic Hierarchy Process. *Research in Higher Education*, 38, 593–614. https://doi.org/10.1023/A:1024948630255

Lin, H.F. (2010). An application of fuzzy AHP for evaluating course website quality. *Computers & Education*, *54*, 877-888. https://doi.org/10.1016/j.compedu.2009.09.017

Mehregan, M.R., Jamporazmey, M., Hosseinzadeh, M., & Mehrafrouz, M. (2011a). Proposing an approach for evaluating e-learning by integrating critical success factor and fuzzy AHP. *International Conference on Innovation, Management and Service, Singapore.*

Mehregan, M.R., Jamporazmey, M., Hosseinzadeh, M., & Mehrafrouz, M. (2011b). Application of fuzzy analytic hierarchy process in ranking modern educational systems' success criteria. *International Journal of e-Education*, *1*(4), 299-304.

Melon, M.G., Beltran, P.A., & Cruz, M.C.G. (2008). An AHP-based evaluation procedure for Innovative Educational Projects: A face-to-face vs. computer-mediated case study. *Omega*, *36*, 754-765. https://doi.org/10.1016/j.omega.2006.01.005

Mendoza, G.A., Prabhub, R. (2000). Multiple criteria decision-making approaches to assessing forest sustainability using criteria and indicators: A case study. *Forest Ecology and Management*, *131*, 107-126.

Mohammed, H.J., Kasim, M.M., & Shaharanee, I.N. (2018). Evaluating of e-learning approaches using AHP-TOPSIS technique, *Journal of Telecommunication, Electronic and Computer Engineering*, *10*, 1-10.

Mondal, K., & Pramanik, S. (2014). *Neutrosophic Sets and Systems*, *6*, 28-34.

Murakoshi H., Kawarasaki T., & Ochimizu K. (2001). Comparison using AHP Web-based learning with classroom learning**,** *Proceedings of Symposium on Applications and the Internet Workshops***,** 67-73. https://doi.org/10.1109/SAINTW.2001.998212

Mustaffa, W.S.W., Shokory, S.M., & Kamis, H. (2006). The Analytical Hierarchy Process: Multi-Criteria Decision Making for Promoting Academic Staff in Higher Education. *The Journal of Global Business Management*, *2*(2).

Nagpal, R., Mehrotra, D., Sharma, A., & Bhatia, P. (2013). ANFIS method for usability assessment of the website of an educational institute. *World Applied Sciences Journal*, *23*(11), 1489–1498. https://doi.org/10.5829/ idosi.wasj.2013.23.11.790

Nagpal, R., Mehrotra, D., Bhatia, P.K., & Sharma, A. (2015). FAHP approach to rank educational websites on usability. *International Journal of Computing and Digital Systems*, *4*(4), 251–260. http://dx.doi.org/10.12785/IJCDS/040404

Naveed, Q.N., Qureshi, M.R.N., Alsayed, A.O., Muhammad, A., Sanober, S. & Shah, A. (2017). Prioritizing barriers of E-learning for effective teaching-learning using fuzzy analytic hierarchy process (FAHP*). 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 1-8.

Naveed, Q.N., Qureshi, M.R., Tairan, N., Mohammad, A., & Shaikh, A. (2020). Evaluating critical success factors in implementing e-learning system using multi-criteria decision-making. *PLoS ONE*, *15*(5), https://doi.org/10.1371/journal.pone.0231465

Nilashi, M., & Janahmadi, N. (2012). Assessing and prioritizing affecting factors in e-learning websites using the AHP method and fuzzy approach. *Information and Knowledge Management*, *2*(1), 46-61.

Nikoomaram, H., Mohammadi, M., Javad Taghipouria, M., & Taghipourian, Y. (2009). Training performance evaluation of administration sciences instructors by fuzzy MCDM approach. *Contemporary Engineering Sciences*, *2*(12), 559–575.

Omurbek, N., Karaatli, M., & Yetim, T. (2014). Analitik hiyerarsi surecine dayali TOPSIS ve VIKOR yöntemleri ile ADIM universitelerinin değerlendirilmesi [Evaluation of ADIM universities with TOPSIS and VIKOR methods based on analytical hierarchy process]. *Selcuk Universitesi Sosyal Bilimler Dergisi*, Dr.Mehmet YILDIZ special issues. 189-207.

Opricovic, S., (1998). *Multicriteria Optimization of Civil Engineering Systems* [Doctoral dissertation, Faculty of Civil Engineering].

Ozdemir, M.S., & Gasimov R.N. (2004). The analytic hierarchy process and multiobjective 0-1 faculty course assignment. *European Journal of Operational Research*, *157*, 398-408. https://doi.org/10.1016/S0377-2217(03)00189-9

Ozkul, A.E., Girginer, N., & Ozturk, Z.K. (2007*). Multi-Criteria Evaluation of Distance Education Implementation Models using Analytic Hierarchy Process*, Proceedings of the 21st Annual Conference Empowering Asia through Partnership in Open and Distance Learning, 87.

Ozturk, Z.K. (2014). Using a multi-criteria decision making approach for open and distance learning system selection. *Anadolu University Journal of Science and Technology– An Applied Sciences and Engineering*, *15*(1), 1-14.

Paksoy, S. (2015). Ülke göstergelerinin vikor yöntemi ile değerlendirilmesi [Evaluation of Country Indicators by Vikor Method]. *Ekonomik ve Sosyal Araştırmalar Dergisi*, *11*(2), 153-169.

Perez Vergara, I.G., Arias Sa´nchez, J.A., Poveda-Bautista, R., & Diego-Mas J.A. (2020). Improving distributed decision making in inventory management: A combined ABC-AHP approach supported by teamwork. *Complexity*, 3–5, 1–13.

Politis, Y., & Siskos, Y. (2004). Multicriteria methodology for the evaluation of a Greek engineering department. *European Journal of Operational Research*, *156*, 223-240. https://doi.org/10.1016/S0377-2217(02)00902-5

Ray, S. (2007). Selecting a doctoral dissertation supervisor: analytical hierarchy approach to the multiple criteria problem. *International Journal of Doctoral Studies*, *2*, 23-32. https://doi.org/10.28945/55

Roy, B. (2005). Paradigms and challenges. In J. Figueira, S. Greco, & M. Ehrgott (Eds.), *Multiple criteria decision analysis: State-of-the-art surveys* (pp. 3–24). Springer.

Saaty, T. L. (1999). Basic theory of the analytic hierarchy process: How to make a decision. *Revista de la Real Academia de Ciencias Exactas Fisicas y Naturales*, 93(4), 395-423.

Saaty, T.L., & Ramanujam, V. (1983). An objective approach to faculty promotion and tenure by the analytic hierarchy process. *Research in High Education*, *18*, 311-331. https://doi.org/10.1007/BF00979603

Sagir, M., & Ozturk, Z.K. (2010). Exam scheduling: Mathematical modeling and parameter estimation with the Analytic Network Process approach. *Mathematical and Computer Modelling*, *52*, 930-941. https://doi.org/10.1016/j.jallcom.2011.02.170

Shee, D.Y., & Wang, Y.S. (2008). Multi-criteria evaluation of the web-based e-learning system: A methodology based on learner satisfaction and its applications. *Computer & Education*, *50*. 894-905. https://doi.org/10.1016/j.compedu.2006.09.005

Soba, M. (2012). Universite öğrencilerinin performanslarinin akademisyenler tarafından analitik hiyerarşi sureci ile değerlendirilmesi [Evaluation of university students' performances by academics through the analytical hierarchy process]. *Electronic Journal of Social Sciences*, *11*(42), 368-381.

Subbaiah, K.V., Shekhar, N.C., & Kandukuri, N.R. (2014). Integrated DEA/TOPSIS approach for the evaluation and ranking of engineering education institutions-a case study. *International Journal of Management Science and Engineering Management*, *9*(4), 249-264. https://doi.org/10.1080/17509653.2014.902758

Syamsuddin, I. (2012). Fuzzy multi-criteria evaluation framework for E-learning software quality. *Academic Research International*, *2*(1), 139-147.

Tekindal, B., & Erumit, A.K. (2007). Analitik hiyerarşi sureci (AHS) ve bulanık AHS yöntemlerinin yüksek lisans öğrencisi seçimi problemi üzerinde karşılaştırılması [Comparison of analytical hierarchy process (AHS) and fuzzy AHP methods on graduate student selection problem]. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi*, *21*, 14-37.

Tezergil, S. (2016). Vikor yöntemi ile Türk bankacılık sektörünün performans analizi [Evaluation of Country Indicators by Vikor Method]. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *38*(1), 357-373. https://doi.org/10.14780/iibd.92056

Timor, M., (2011). *Analitik hiyerarşi prosesi* [*Analytical hierarchy process*]. Türkmen Kitabevi.

Turki, A., & Duffuaa, S. (2003). Performance measures for academic departments. *International Journal of Educational Management*, *17*(7), 330-338. https://doi.org/10.1108/IJEM-09-2014-0129

Tzeng, G.H., Chiang, C.H., & Li, C.W. (2007). Evaluating intertwined effects in learning programs: A novel hybrid MCDM model based on factor analysis and DEMATEL. *Expert Systems with Applications*, *32*, 1028-1044. https://doi.org/10.1016/j.eswa.2006.02.004

Wu, H.Y., Chen, J.K., Chen, I.S., & Zhuo, H.H. (2012). Ranking universities based on performance evaluation by a hybrid MCDM model. *Measurement*, *45*, 856-880. https://doi.org/10.1016/j.measurement.2012.02.009

Yazdani, B.O., Yaghoubi, E.S., & Giri, E.S. (2011). Factors affecting the empowerment of employees (an empirical study). *European Journal of Social Sciences*, *20*(2), 267-274.

Yigit, T., Isik, A.H., & Ince, M. (2014). Web-based learning object selection software using analytical hierarchy process. *IET Software*, *8*(4), 174–183. https://doi.org/10.1049/iet-sen.2013.0116

Zare, M., Pahl, C., Rahnama, H., Nilashi, M., Mardani, A., Ibrahim, O., & Ahmadi, H. (2016). Multi-criteria decision-making approach in e-learning: A systematic review and classification. *Applied Soft Computing*, *45*, 108-128. https://doi.org/10.1016/j.asoc.2016.04.020

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 29–55

https://doi.org/10.21449/ijate.995005

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

Reliability and validity of the Turkish version of the teachers' basic ICT competence beliefs scale

Pinar Korukluoglu[1,*], Bulent Alci[2], Charlott Rubach[3]

[1]Yildiz Technical University, Faculty of Education, Department of Educational Sciences, Istanbul, Türkiye
[2]Yildiz Technical University, Faculty of Education, Department of Educational Sciences, Istanbul, Türkiye
[3]University of Rostock, Philosophical Faculty, Institute for School Pedagogy and Educational Research (ISB), Rostock, Germany

ARTICLE HISTORY

Received: Sep. 13, 2021

Accepted: Jan. 15, 2023

Keywords:

Cross-cultural validation,
Pedagogical issues,
ICT competence beliefs,
Teacher professional development.

Abstract: The present study seeks to adapt the Teachers' Basic Information Communication Technology (ICT) Competence Beliefs Scale, developed by Rubach and Lazarides (2021), into Turkish and test the adapted scale's validity and reliability. The initial step involved conducting a linguistic equivalence of the scale from English to Turkish with 62 English language teachers in a pre-test. Subsequently, the Turkish version of the scale was administered to 356 teachers (69.7% female, 30.3% male) in Turkey to assess its validity and reliability. Participating teachers were from different subjects (e.g., 9.8% science, 7.9% mathematics, 3.7% social science) and school types (27.5% primary school, 55.3% secondary school, 17.1% others). Results of confirmatory factor analysis indicated the original six-factor structure with three first-order and three second-order factors that best fitted the data. The same competence dimensions were indicated in the Turkish contexts as in the original instrument, i.e., information and data literacy; communication and collaboration; digital content creation; safety and security; problem-solving; analyzing and reflecting. The correlations between all six first-order factors were between $.58 \geq r \geq .79$. All factors showed good reliability indices, i.e., $\alpha > .83$, $\omega > .83$ and CR > .72. The adapted instrument was found to be invariant across gender. Mean-level differences among gender groups point to one difference with male teachers reporting higher competence beliefs for digital content creation compared to female teachers. In conclusion, the results of this replication study support the cross-cultural transferability of the original Teachers' Basic ICT Competence Beliefs instrument developed by Rubach and Lazarides (2019).

1. INTRODUCTION

The competence to use Information and Communication Technologies (ICT) is widely recognized as a crucial skill in the current era (Ferrari, 2013; OECD, 2018; Voogt & Roblin, 2012; Wang, Sigerson, & Cheng, 2019). With the rapid advancement of technology in recent decades, society has transformed from an industrial-based society to a digital information society (Anderson, Van Weert, & Duchâteau, 2002; Bayazıt & Seferoğlu, 2009; Parlak, 2017).

*CONTACT: Pinar Korukluoglu ✉ pinarkoruklu16@gmail.com Yildiz Technical University, Faculty of Education, Department of Educational Sciences, Istanbul, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

As a result, the educational sector has also been influenced by technological advancements and ICT is seen as a means to further develop, enhance and innovate the learning processes (Kocaman Karoğlu, Bal, & Çimşir, 2020; Parlak, 2017; Redecker & Punie, 2017; Voogt & Roblin, 2012). In response to these changes, ICT has been integrated into educational systems as a crucial learning tool, and the infrastructure of information and communication technologies has been developed in various countries, such as the "2.0 School Program" in Spain, the "Digital School Plan" in Hungary, the "Smart School Program" in Italy (Gil-Flores, Rodríguez-Santero, & Torres-Gordillo, 2017), and in Turkey, the "Education and Information Network (EBA)" and the "Teacher Information Network (ÖBA)" (EBA, 2020; İzmirli, 2015; ÖBA, 2022). Additionally, new technologies such as artificial intelligence and augmented reality have been utilized to support e-learning and digital-based education (Kapur et al., 2018; Kocaman Karoğlu, Bal, & Çimşir, 2020). The integration of technology in education has the potential to improve educational processes and increase learning efficiency, with a focus on students' future professional education and life skills (Seufert, Guggemos, & Sailer, 2021).

The digital transformation and digitalization of education also bring new responsibilities for teachers, including the mastery of digital tools to enhance their teaching and to facilitate their students' ICT competence (Redecker & Punie, 2017; Rubach & Lazarides, 2019; Şad & Nalçacı, 2015; Yurdakul, Dönmez, Altınok, & Odabaşı, 2013). This has given rise to the concept of digital leadership, which requires the adoption and utilization of new technology, the creation and management of technology-related jobs, and the motivation of individuals to achieve their goals in the digital space in order to transform schools into learning spaces suited for the digital age (Asri & Darma, 202; Zhong, 2017). As a result, teachers must be competent in using ICT and fulfil their digital leadership role (Eickelmann & Vennemann, 2017; Hatlevik, Throndsen, Loi, & Gudmundsdottir, 2018). Several frameworks have been introduced to define the basic competencies that teachers should possess in order to fulfil their professional responsibilities. The Technological Pedagogical Content Knowledge (TPACK) model suggests that the best implementation of ICT in the learning and teaching process is achieved through the convergence of technological knowledge, along with pedagogical knowledge, and content knowledge (Mishra & Koehler, 2006; Tondeur, Aesaert, Prestridge, & Consuegra, 2018). The TPACK model is composed of three main components: *technology knowledge, content knowledge, and pedagogy knowledge*, and four sub-components: *technological pedagogical knowledge, technological content knowledge, pedagogical content knowledge,* and *technological pedagogical content knowledge* (Koehler & Mishra, 2005).

Another theoretical approach, as described by Krumsvik (2014) and Rubach and Lazarides (2021), differentiates teachers' ICT competencies into two categories: basic and pedagogical. With regards to educational policy, the Information and Communication Technology Competency Framework for Teachers (ICT-CFT) has been established by the International Society for Technology in Education (ISTE, 2008) and UNESCO (2011). The basic ICT competencies of teachers are categorized as professional competencies, including critical thinking skills, generic skills, ICT skills for professional development, decision-making skills, change management skills, cooperative working skills, and effective communication skills (Anderson, Van Weert, & Duchâteau, 2002; UNESCO, 2011).

Concentrating on the pedagogical ICT competencies of teachers, the International Society for Technology in Education (ISTE, 2008) categorizes these competencies as the orchestration of seven dimensions relevant to teaching and student support. These dimensions include the ability to discover technological innovations for student development, serve as a digital education leader, support students in realizing their responsibilities in the digital world and making positive contributions, collaborate with students and colleagues to use digital resources, create innovative digital learning environments considering individual student differences, facilitate learning with technology, and analyze data to assist students in reaching their learning goals as an instructional leader. In Turkey, ICT competencies are deemed mandatory for teachers' generic competencies, as per the Ministry of National Education (MoNE, 2006; MoNE, 2017).

The requirement to establish training programs that aim to improve teachers' ICT competencies and to evaluate the effectiveness of these programs is becoming increasingly crucial (Ananiadou & Claro, 2009; Ferrari, 2012; Ilomäki, Paavola, Lakkala et al., 2016; ISTE, 2008; Kultusministerkonferenz, 2016; OECD, 2018; UNESCO, 2011). To evaluate and enhance these programs, the development of valid and reliable evaluation tools to assess teachers' ICT competencies and related competence beliefs is necessary.

In current educational research, instruments aimed at evaluating teachers' competence beliefs, specifically their perceived ICT competencies, have primarily been utilized (Gerick, Eickelmann, & Bos, 2017; Tondeur, Braak, & Valcke, 2007; Tondeur, Aesaert, Prestridge, & Consuegra, 2018). Competence belief has been defined as individuals' assessments of their competencies in various areas (Muenks, Wigfield, & Eccles, 2018). Different theoretical frameworks have differentiated competence beliefs, including specific concepts such as achievement-related expectancies for success (Eccles et al., 1983) and self-efficacy (Bandura, 1977). The underlying theoretical assumption is that competence beliefs, competencies, and related motivational beliefs, such as subjective task values, have an impact on teachers' utilization of ICT in the classroom. Research has shown that basic ICT competence beliefs have a predictive effect on teachers' utilization of ICT, particularly for innovative instruction, whereas pedagogical ICT competence beliefs significantly impact teachers' teaching quality and their ability to incorporate ICT content into their teaching (Angelie & Valanides, 2009; Guggemos & Seufert, 2021; Hatlevik, 2017).

Numerous studies have noted the disparity in the characterization of ICT competencies across various frameworks (Fraillon et al., 2014; Koh et al., 2013; Scherer et al., 2017; Vanderlinde & Van Braak, 2010). The European Digital Competences Framework (Digcomp-Ferrari, 2012) differentiated ICT competencies into six dimensions, namely: information and data literacy, communication and collaboration, digital content creation, safety and security, problem-solving, and analysis and reflection. Furthermore, it introduced a pedagogical ICT license aimed at enhancing teachers' pedagogical competencies. In a recent study, Rubach and Lazarides (2021) developed and validated a scale to assess teachers' basic ICT competence beliefs across various competence dimensions. The scale was designed based on the European Digital Competence Framework (Ferrari, 2012) and the German educational policy framework (Kultusministerkonferenz, 2016) and consisted of six factors that capture the competence dimensions described in previous studies (Rubach & Lazarides, 2021; Ferrari, 2013). These factors include *information and data literacy* (second-order factors: searching, storing and organization), *communication and collaboration*, *digital content creation, safety and security, problem-solving* (second-order factors: operation and usage, comprehension and development), and *analysis and reflection* (second-order factors: analysis of distribution and risk, analysis of business activities).

Despite the emphasis placed on ICT competencies for teachers by the Ministry of National Education (MoNE) in Turkey, a valid and reliable instrument is still needed that can be used to investigate all assumed dimensions of teachers' basic ICT competence beliefs in the Turkish context. Previous instruments used in Turkey to assess ICT competencies have limitations, such as being primarily designed for pre-service teachers and focusing only on the level of ICT usage rather than competence beliefs (Anagün et al., 2016; Gökçearslan et al., 2019; Kutluca et al., 2010; Türel et al., 2017). Moreover, instruments guided by the Technological Pedagogical Content Knowledge-Practice (TPACK Pratik) model tend to only measure the use of ICT with more general TPACK features (Ay et al., 2015). Therefore, there is a need for an instrument that measures the full range of relevant ICT skills required for competent usage.

The instrument developed by Rubach & Lazarides (2019) addresses this need but was developed for the German context. Thus, this study aimed to validate the instrument for the Turkish context. The instrument developed by Rubach and Lazarides (2021) is deemed appropriate for validation in Turkey for several reasons, including its emphasis on the necessary items for the competence beliefs of in-service teachers and the factors and sub-factors were created in alignment with current, need-oriented comprehensive scientific research (Ferrari, 2013). Furthermore, the adaptation of this instrument to the Turkish context and investigation of its validity and reliability is expected to contribute to the professional development of both teacher candidates and working teachers in Turkey, as it will provide a means of identifying ICT training needs in the 21st century that meet international criteria (Ferrari, 2013). Thus, we assume the same proposed structure as in Rubach and Lazarides (2021).

The significance of the examination of ICT competence in teachers is widely acknowledged on a transnational level, as it is considered to be a crucial component of effective teaching practices in the 21st century (Parlak, 2017; Palvia et al., 2018). 21st century ICT competence of teachers is indispensable in creating an effective teaching environment (Fraillon et al., 2014). This viewpoint is supported by the Ministry of National Education (MoNE) in Turkey, which recognizes the importance of ICT competencies in teacher training and professional development (MoNE, 2006; MoNE, 2017). Professional development training and its evaluation are needed for teachers beyond country borders to increase their competency by adopting ICT in the classroom (Galanouli et al., 2004). Thus, it is helpful to use the same instrument to compare the motivational traits of teachers across countries. Hence, scale adaptation studies in this subject are essential for repeating and comparing cross-cultural studies. Ensuring the scales' validity in different cultures makes it possible to prepare international education programs.

In light of these considerations, this study aims to adapt the "Teachers' Basic ICT Competence Beliefs" instrument developed by Rubach and Lazarides (2021) into Turkish and test its validity and reliability in the Turkish context. The following research questions guided the study:

RQ 1: Is the Turkish version of the "Teachers' Basic ICT Competence Beliefs" instrument valid?
RQ 2: Is the Turkish version of the "Teachers' Basic ICT Competence Beliefs" instrument reliable?

2. METHOD

2.1. Participants

The sample for the study was drawn from the central districts of Bursa, Turkey and was obtained through the method of convenience sampling, which is a type of purposive sampling. This method was chosen as it allows for the acquisition of relevant data in a timely manner (Patton, 2018). The sample for the pretest consisted of 62 English Language Teachers, with 58.1% of the participants being female and 41.9% being male. A demographic analysis of the pre-test participants is presented in Table 1, which indicates that 12.9% of the teachers were under 26 years of age, 22.6% were between 26-34 years old, 45.2% were between 35-44 years old, and 19.4% were between 45-54 years old.

Table 1. *Demographic Information of the Pre-test Participants.*

		n	%
Gender	Female	36	58.1
	Male	26	41.9
Age	25 and under	8	12.9
	26-34	14	22.6
	35-44	28	45.2
	45-54	12	19.4
Subject	English language	62	100
Total		62	100

Table 2. *Demographic Information of the Main Study Participants.*

		n	%
Gender	Female	248	69.7
	Male	108	30.3
Age	25 and under	4	1.1
	26-34	96	27.0
	35-44	145	40.7
	45-54	89	25.0
	55 and above	22	6.2
Subject	Pre-school Teachers	25	7.0
	Primary School Teachers	82	23.0
	Turkish Language	37	10.4
	Mathematics	28	7.9
	Science	35	9.8
	Social Science	13	3.7
	English Language	28	7.9
	Visual Art	9	2.5
	Technology and Design	8	2.2
	Physical Education	15	4.2
	Religious Culture and Ethics Mu-	20	5.6
	sic Teacher	10	2.8
	School Guidance Counselors	22	6.2
	Information Technology	7	2.0
	Philosophy	5	1.4
	History	5	1.4
	Literature	3	.8
	Vocational Training Teachers	4	1.1
School type	Pre-school	24	6.7
	Primary school	98	27.5
	Secondary school	157	55.3
	High school	37	10.4
TOTAL		356	100

The sample for the main study consisted of 356 teachers, with 69.7% being female and 30.3% being male. The sample size of 356 participants was deemed sufficient for conducting factor analysis in the scale adaptation study. Field (2018) and Tabachnick and Fidell (2013), emphasized that the sample size for such studies should be at least 300 cases in order to ensure the reliability of the instruments. The demographic characteristics of the sample are detailed in Table 2, which highlights the age distribution of the participants, with 4 (1.1%) being less than 25 years old, 96 (27%) being between 26 and 34 years old, 145 (40.7%) being between 35 and 44 years old, 89 (25%) being between 45 and 54 years old, and 22 (6.2%) being older than 55 years old. In terms of their teaching roles, 6.7% of the participants were pre-school teachers, 27.5% were primary school teachers, 55.3% were secondary school teachers, and 10.4% were high school teachers.

2.2. Instruments

The "Teachers' Basic ICT Competence Beliefs" instrument was developed by Rubach and Lazarides (2021) and consists of 32 items divided into six competence domains: *information data literacy* (6 items), *communication and collaboration* (6 items), *digital content creation* (4 items), *safety and security* (4 items), *problem-solving* (7 items), and *analyzing and reflecting* (5 items). Three of the competence domains possess a second-order structure: *information data literacy* (searching, storing, and organization), *problem-solving* (operation and usage, comprehension and development), and *analyzing and reflecting* (analysis of distribution and risk, analysis of business activities). The data fit was analyzed using statistical indices such as the Kaiser-Meyer-Olkin coefficient (KMO = .93) and Bartlett's Test of Sphericity (x^2 [1378] = 9290.98, p <.0001) in the exploratory factor analysis (EFA), and x^2/df= 1.48 [654.73/441], CFI = .96, RMSEA= .04 in the confirmatory factor analysis (CFA) which indicated a good fit. The reliability of the instrument was supported with values of McDonald’s omega (ω) ranging between $.63 \geq \omega \geq .93$. The original scale utilized the five-point Likert type, ranging from 1 (strongly disagree) to 5 (strongly agree), with no reverse items. In this study, the 32 items were translated into Turkish and the same Likert scale was used as in the original instrument.

2.3. Procedure and Data Analysis

The translation of the "Teachers' Basic ICT Competence Beliefs" instrument from English to Turkish was carried out using a forward-backward translation technique. Initially, three English language teachers in Turkey were tasked with translating the English version of the instrument into Turkish. These teachers then collaborated to reconcile any differences in their translations and arrived at a consensus for the final version of the Turkish translation. The final version of the Turkish instrument was reviewed for linguistic and cultural appropriateness by an expert in linguistics who is proficient in both English and Turkish. The Turkish version of the instrument was then back-translated into English by two academics working at a university's English preparatory school, and the two back-translations were compared for word compatibility and cultural-linguistic equivalence.

In the pre-test phase, the equivalence of the original and translated versions of the scale was assessed through the completion of both the English and Turkish versions of the instrument by teachers. Correlation coefficients between the original and the translated versions of the scale and paired-samples t-test were analyzed. For the main study, a second sample of teachers was recruited to evaluate the validity and reliability of the Turkish version of the instrument. The analysis was performed using various software and techniques. Using SPSS 23.0 and SPSS AMOS 26.0, Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were conducted as well as reliability indices Cronbach Alpha (α), McDonald’s (ω), Mplus 8.1 were utilized to test measurement invariance across gender groups and Microsoft Excel was utilized to calculate the Composite Reliability (CR) coefficient. The study adhered to the

"Guidelines for Translating and Adapting Tests" (IJATE, 2014).

The study utilized Exploratory Factor Analysis (EFA) to examine the number of factors and factor loadings of the items in the scale and their relationships. The sample size was analyzed using the Kaiser Meyer Olkin (KMO= .95) coefficient, and the data for factor analysis was analyzed using Barlett's Sphericity test value (χ^2= 10052.01, *df*= 406, $p \leq .001$) with maximum-likelihood estimation and a normal covariance matrix. Factor loadings and variances were used to assess the appropriateness of factors and items, and multicollinearity between factors was examined based on factor correlation, the Variance Magnification Factor (VIF), and tolerance values. The normality assumption of the data was indicated by examining skewness and kurtosis.

Confirmatory Factor Analysis (CFA) was employed to verify the appropriateness of the original instrument's structure after translation and adaptation to a different language and culture (Seçer, 2018; Tabachnick & Fidell, 2013). The fit of the model to the data indicated by CFA was evaluated using various fit indices (Hu & Bentler, 1999), including Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Incremental Fit Index (IFI), and Root Mean Square Error of Approximation (RMSEA). The study also calculated reliability indices such as Cronbach Alpha (α), McDanold's omega (ω), and Composite Reliability (CR). In order for the scale to have qualities such as validity and reliability, it is considered appropriate to test the quality of each item of the scale with item analysis (Tekindal, 2015). Thus, item analyses were carried out to estimate item-total correlation values; the difference between the mean scores of the lower 27% and upper 27% groups of the total scores of the scale was examined with independent t-tests.

The study also tested the measurement invariance across gender groups using Mplus 8.1 (Muthén & Muthén, 1998-2016). We conducted analyses of the measurement invariance as a robustness check to replicate finding on gender differences and similarities reported by Rubach & Lazarides (2021). In order to examine the robustness of the instrument and to replicate the findings of gender differences and similarities reported by Rubach & Lazarides (2021), a measurement invariance analysis was conducted. The configural, metric, and scalar invariance were examined by systematically constraining the factor loadings and item intercepts to equality across males and females. Testing measurement invariance enables to determine similarities and differences across groups and thus tests the robustness of the instrument, e.g., across groups or time. That is, the measurement invariance tests indicated if the expected scores of individuals were independent of group membership or time (Chen, 2007; Wicherts, 2007). Cut-off values for sample sizes n > 300 were used to indicate insignificant changes in the more restrictive model: ΔCFI $\leq$ - .010 and ΔRMSEA $\leq$ .015, or ΔSRMR $\leq$ .030 for step 1 (configural invariance) and values of ΔCFI $\leq$ - 0.010 and ΔRMSEA $\leq$ 0.015 or ΔSRMR $\leq$ .010 for step 2 (metric and scalar invariance) (Chen, 2007).

2.4. Ethical Considerations

As scientific professionals, it is incumbent upon us to ensure the accuracy and reliability of the information we generate and disseminate for the betterment of society. To this end, it is imperative that we adhere to established ethical principles throughout all stages of the scientific research process (TÜBA, 2008). This study was undertaken with due regard for ethical considerations, starting with obtaining permission from the owner of the measurement instrument in accordance with scientific ethical guidelines. Participants in the study were provided with an informed consent form, and their participation was strictly voluntary. No personal information was solicited through the instrument, and the data collected was solely intended for scientific purposes. The analysis, interpretation, and reporting of these data were guided by ethical principles, and the study was approved by the Yildiz Technical University Humanities and Social Sciences Research Academic Ethics Committee (Approval no: 2021/01, dated 21.03.2021) prior to its implementation.

3. RESULTS

3.1. Linguistic Equivalence

The linguistic equivalence stage of this study involved administering both the English and Turkish versions of the scale as an online form at one-week intervals. This methodology is in line with previous studies (Baş & Balaman, 2021; Dündar et al., 2008; Kılıç & Alcı, 2022) which have also employed the application of the original scale and its target language equivalent to a sample group of proficient bilinguals at one-week intervals. It was seen that approximately 30 bilinguals were employed in the studies indicated for this stage. In this study, a sample of 62 participants was recruited for the linguistic equivalence assessment, yielding a sufficient sample size. The associations between the total scores of the scale and the total scores of its factors and second-order factors were then investigated for both the Turkish and English versions, as shown in Table 3. The correlation coefficients (r) between the scores were found to be greater than .84, indicating strong correlations (Büyüköztürk, 2011). Based on these findings, it can be concluded that linguistic equivalence was achieved between the English and Turkish versions of the scale.

Table 3. *Correlation Coefficient between Turkish and English Versions.*

Factors/Second-order factors	*r*
Factor 1: Information and data literacy	.92**
Second-order factor (Factor 1.1): Searching	.91**
Second-order factor (Factor 1.2): Storing and organizing	.88**
Factor 2: Communication and collaboration	.89**
Factor 3: Digital content creation	.88**
Factor 4: Safety and security	.90**
Factor 5: Problem-solving	.95**
Second-order factor (Factor 5.1): Operation and usage	.88**
Second-order factor (Factor 5.2): Comprehension and development	.94**
Factor 6: Analyzing and reflecting	.88**
Second-order factor (Factor 6.1): Analysis of distribution and risks	.84**
Second-order factor (Factor 6.2): Analysis of business activities	.85**
TOTAL	.97**

Note. ***p* < .001

Results on mean differences of factors and second-order factors between the Turkish and English versions are presented in Table 4.

Table 4. *Paired-samples t-test Values between Turkish and English Versions.*

	Language	N	*SS*	X	*t*-test	
					t	*p*
TOTAL	English	62	17.75	126.09	-1.06	.28
	Turkish	62	20.23	126.77		

Note. *p*>.05

The results showed no significant difference between the two versions of the scale. In addition, inter-factors correlation coefficients of the Turkish and English versions are shown in Table 5.

Table 5. *Inter-factor Correlation Coefficients of Turkish Version and English Version.*

	Factor 1		Factor 2		Factor 3		Factor 4		Factor 5		Factor 6	
	Turk.	Eng.	Turk.	Eng.	Turk.	Eng.	Turk.	Eng.	Turk.	Eng.	Turk.	Eng.
Factor 1	1	1	.74**	.58**	.64**	.52**	.69**	.56**	.68**	.60**	.74**	.63**
Factor 2			1	1	.82**	.75**	.71**	.51**	.76**	.61**	.75**	.60**
Factor 3					1	1	.58**	.49**	.83**	.75**	.67**	.53**
Factor 4							1	1	.68**	.55**	.68**	.53**
Factor 5									1	1	.81**	.69**
Factor 6											1	1

Note. **p<.001, Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

Table 5 reveals the absence of significant differences in the correlation values between the two scale factors, thereby providing evidence for the reliability of the Turkish translation of the scale.

3.2. Validity Study

3.2.1. *Exploratory Factor Analysis (EFA)*

This study employed Exploratory Factor Analysis (EFA) to evaluate the structural validity of the scale. The results of the Kaiser Meyer Olkin coefficient indicated that the sample size was adequate (KMO=.95 > .70); The Barlett's test of Sphericity (χ^2 = 10052.01 > .5; df= 406; $p \leq$ .001) confirmed the suitability of the data for factor analysis (Hutcheson & Sofroniou, 1999). In addition, our results were similar to the EFA results of the original scale (KMO = .93; Bartlett's Test of Sphericity= x^2 [1378] =9290.98, $p \leq .001$) (Rubach & Lazarides, 2021).

The EFA, performed using oblique rotation on all 32 items, revealed that three items were double-loaded (item3 in factor1, item12 in factor2, item13 in factor 3; see Table 6). The oblique rotation method rotates factors independently, which does not alter the ratio of total variance explained by the factors (Tabachnick & Fidell, 2013). Consequently, these three items were removed from the scale to avoid overlap (Seçer, 2018). This outcome may be due to differences in understanding or attitudes among teachers in the sample group (Buabeng-Andoh, 2012), or to intercultural differences in individual responses to these items (Ay et al., 2015).

Table 6. *Excluded Items.*

Excluded item no	Factor No	Excluded items (Original version)	Excluded items (Turkish Version)
Item 3	Factor 1.1	I am critical about information, sources and data in digital environments	Dijital ortamdaki bilgi, veri ve kaynaklar konusunda eleştirel bir yapıdayım.
Item 12	Factor 2	I can share my experiences with digital media in interactions with others	Dijital medya ile ilgili deneyimlerimi, başkalarıyla etkileşim halinde paylaşabilirim.
Item 13	Factor 3	I can use familiar apps and programs according to my needs.	İhtiyaçlarım doğrultusunda, aşina olduğum uygulama ve programları kullanabilirim

Finally, a six-factor solution, consisting of three first-order factors with two second-order factors each (29 items), was subjected to analysis (as depicted in Table 7). The factor loadings range between .46 ≥ λ ≥ .93, with a loading value of λ ≥ .45 considered as appropriate, and a threshold value of .30 considered acceptable in some cases (Büyüköztürk, 2011; Tabachnick &

Fidell, 2013). The common variance values of the factors, as specified in Table 7, show that the variance of the factors ranges from .75 to .92, with a factor variance above .66 considered a proper solution (Büyüköztürk, 2011; Tavşancıl, 2014). Additionally, an explained variance of .30 or above is considered adequate for scales with one factor, while a higher explained variance is expected for scales with multiple factors (Büyüköztürk, 2011; Çokluk et al., 2010; Tabachnick & Fidell, 2013). The explained variance of the scale in this study is 83.89%, suggesting a sound structure.

Table 7. *Factor Loadings and Factor Variance.*

	Factor Loadings									
Items	Factor 1.1	Factor 1.2	Factor 2	Factor 3	Factor 4	Factor 5.1	Factor 5.2	Factor 6.1	Factor 6.2	Factor Variance
1	.90									.87
2	.79									.86
3		.93								.76
4		.83								.90
5		.73								.87
6			.65							.80
7			.63							.82
8			.59							.82
9			.58							.77
10			.48							.75
11				.90						.88
12				.86						.92
13				.84						.90
14					.83					.80
15					.80					.85
16					.70					.77
17					.51					.80
18						.67				.82
19						.60				.81
20						.60				.85
21						.55				.81
22							.78			.82
23							.74			.83
24							.74			.81
25								.53		.83
26								.53		.86
27								.46		.81
28									.85	.91
29									.83	.90

Note. Total Variance Explained: %83.89, Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

The correlation coefficients between the factors were below 0.80, which suggests that no multicollinearity problem was present (Büyüköztürk, 2011). To further verify this, the Variance Magnification Factor (VIF) was calculated. The criteria established by Tabachnick and Fidell (2013) dictate that if the VIF value is higher than 10, there is multicollinearity between variables. Besides this, tolerance values less than 0.10 indicate collinearity (Daoud, 2017). The VIF values for the present study, ranging from 2.483 to 4.937 (VIF<10), suggesting that there was no multicollinearity between the factors (Büyüköztürk, 2011). Moreover, the tolerance values for each factor, ranging from 0.20 to 0.40 (the values > .10), supported this conclusion.

3.2.2. *Normal Distribution Analysis*

The univariate normal distribution of the data was evaluated by means of the skewness and kurtosis values, as proposed by Tabachnick and Fidell (2013). The normality of the data was assessed for each item and factor based on the skewness and kurtosis values, as demonstrated in Table 8. The analysis of normality revealed that the obtained data had a skewness of -.268 and a kurtosis of -.445, which indicated a normal distribution within the bounds of ±3, according to the criteria established by Tabachnick and Fidell (2013) and Trochim and Donnelly (2006). Another cut off to determine substantial non-normality is either an absolute skew value larger than 2 or an absolute kurtosis larger than 7 (Kim, 2013). As reported in Table 8, values for kurtosis and skewness showed normality for 29 items and each factor.

Table 8. *Normality of Data Results.*

Item/Factor	N	Skewness	Kurtosis
Factor 1	356	-.526	-.498
Item 1	356	-.883	-.209
Item 2	356	-.836	.053
Item 3	356	-.568	-.497
Item 4	356	-.589	-.442
Item 5	356	-.507	-.687
Factor 2	356	-.812	.029
Item 6	356	-.991	.292
Item 7	356	-.985	.368
Item 8	356	-.683	-.391
Item 9	356	-.786	-.081
Item 10	356	-.767	-.279
Factor 3	356	-.227	-.888
Item 11	356	-.227	-.937
Item 12	356	-.274	-.969
Item 13	356	-.349	-.820
Factor 4	356	-.693	-.245
Item 14	356	-.920	.271
Item 15	356	-.646	-.404
Item 16	356	-.644	-.389
Item 17	356	-.958	.296
Factor 5	356	-.324	-.253
Item 18	356	-1.250	1.365
Item 19	356	-.913	.330
Item 20	356	-.805	.210
Item 21	356	-.444	-.375
Item 22	356	-.052	-.721
Item 23	356	-.151	-.836
Item 24	356	.347	-.870

Factor 6	356	-.325	-.311
Item 25	356	-.333	-.639
Item 26	356	-.411	-.396
Item 27	356	-.491	-.266
Item 28	356	-.338	-.596
Item 29	356	-.278	-.596
Total	356	-.445	-.268

Note. Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

3.2.3. *Confirmatory Factor Analysis (CFA)*

The Maximum-Likelihood estimation was utilized to test the same Confirmatory Factor Analysis (CFA) model as presented in Rubach and Lazarides (2021). Based on the 32-item CFA, the results χ^2 (438) = 1266.01; CFI = .92, TLI = .91, RMSEA = .073, SRMR = .07. However, examination of the exploratory factor analysis (EFA) indicated the presence of double-loaded items. Consequently, a CFA was conducted using the 29 item solution, which was determined to be an appropriate model with acceptable model fit indices, as shown in Figure 1.

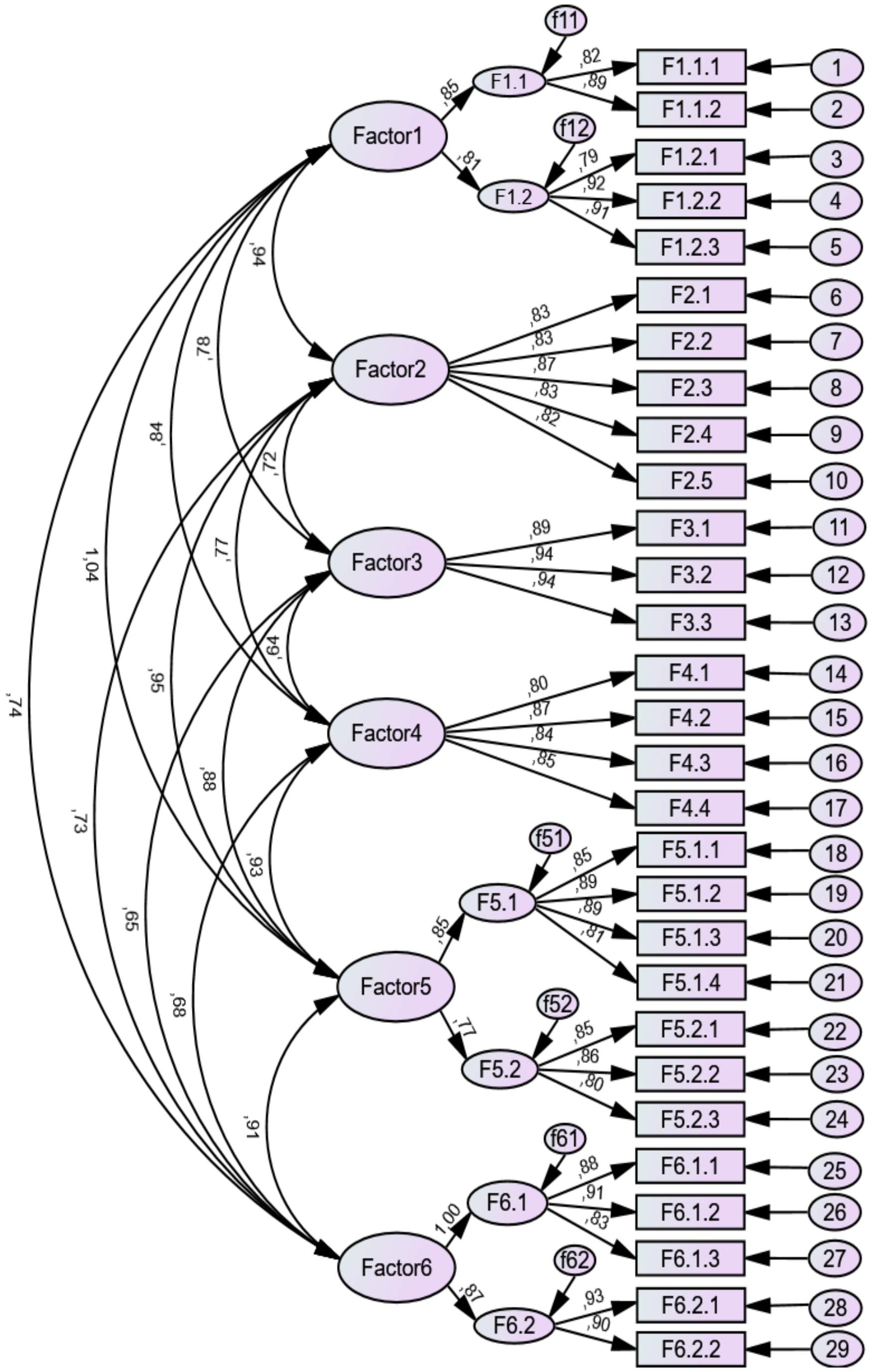
The results of the values obtained from the CFA are revealed in Table 9.

Table 9. *CFA Fit Indices and CFA Results.*

Fit Indices	Perfect Fit	Acceptable Fit	Model fit indices (Rubach & Lazarides)	Model fit indices (Turkish version)
χ^2/df	$0 \leq \chi^2/df \leq 2$	$2 \leq \chi^2/df \leq 3$	1.48	2.96
RMSEA	0≤RMSEA≤ .05	.05 ≤RMSEA≤ .08	.04	.07
CFI	.95 ≤CFI ≤1	.90 ≤CFI ≤ .95	.96	.93
TLI	.95 ≤TLI ≤1	.90 ≤TLI ≤ .95	.95	.92
IFI	.95 ≤IFI ≤1	.90 ≤IFI ≤ .95	--	.93

In Figure 1, it is observed that the factor structure of the Turkish version is consistent with the German version proposed by Rubach & Lazarides (2021). The 29-item solution was found to comprise six second-order factors, which encompass *information and data literacy* (comprising searching, storing and organization), *communication and collaboration*, *digital content creation, safety and security, problem-solving* (encompassing operation and usage, comprehension and development), and *analyzing and reflecting* (encompassing analysis of distribution and risk, analysis of business activities).

Figure 1. *CFA Model.*



3.3. Reliability Study

The examination of internal consistency was performed through the utilization of three measures: the Cronbach Alpha Coefficient (α) (Cronbach, 1951), McDonald's Omega (ω) (McDonald, 1999), and the Composite Reliability Coefficient (CR) (Bacon, Sauer, & Young, 1995) (see Table 10 for further details).

Table 10. *Cronbach Alpha (α), McDonald's omega (ω), Composite Reliability (CR) of the Scale Factors*

	Turkish version				Original version (Rubach & Lazarides, 2021)
	α	ω	Cr	Number of Items	ω
Factor 1	.89	.89	.92	5	-
Second-order factor (Factor 1.1)	.83	.83	-	2	.81
Second-order factor (Factor 1.2)	.90	.91	-	3	.63
Factor 2	.92	.92	.72	5	.86
Factor 3	.94	.94	.90	3	.91
Factor 4	.90	.90	.80	4	.87
Factor 5	.90	.89	.85	7	-
Second-order factor (Factor 5.1)	.91	.91	-	4	.91
Second-order factor (Factor 5.2)	.87	.87	-	3	.85
Factor 6	.93	.92	.78	5	-
Second-order factor (Factor 6.1)	.90	.90	-	3	.86
Second-order factor (Factor 6.2)	.91	.91	-	2	.93
TOTAL	.97	.97	.96	29	

Note. Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

The internal consistency of the data was evaluated using the Cronbach's Alpha Coefficient (α) and McDonald's Omega (ω) and the Composite Reliability (CR) (Cronbach, 1951; McDonald, 1999; Bacon, Sauer, & Young, 1995). According to George & Mallery (2003) and Kılıç (2016), the acceptable range of α is $0.6 \leq \alpha < 0.7$, good range is $0.7 \leq \alpha < 0.9$, and excellent when $\alpha \geq 0.9$. The results showed that the overall Cronbach's Alpha coefficient was .97, indicating excellent reliability, and the Cronbach's Alpha coefficients for all factors were between $.83 \leq \alpha < .94$. The results also indicated excellent reliability for McDonald's Omega with an overall coefficient of .97 and a range between $.83 \leq \omega < .94$. These values are consistent with the findings of Rubach and Lazarides (2021), who reported McDonald's Omega coefficients ranging between $.63 \leq \omega < .93$. The Composite Reliability coefficient, calculated for each factor CR>.72 and the total scale, was found to be reliable with a value of CR = .96, demonstrating structural equality (Bacon, Sauer, & Young, 1995).

Furthermore, inter-factor correlation coefficients (r) were analyzed and presented in Table 11 for both the Turkish and German versions.

Table 11. *Inter-factor Correlation Coefficients between factors for the Turkish Version (before the slash) and the original version by Rubach & Lazarides (2021, behind slash).*

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Factor 1	1/ 1	.79**/.95**	.65**/.68**	.71**/.81**	.76**/.74**	.60**/.71**
Factor 2		1/ 1	.67**/.66**	.70**/.67**	.76**/.66**	.66**/.62**
Factor 3			1/ 1	.58**/.72**	.75**/.88**	.60**/.53**
Factor 4				1/ 1	.74**/.76**	.63**/.70**
Factor 5					1/ 1	.76**/.67**
Factor 6						1/ 1

Note. **p<.001, Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

Table 11 presents the results of the inter-factor correlation analysis, with the coefficients ranging from .58 to .79, which are statistically significant ($.58 < r < .79$; $p < .001$). As per Büyüköztürk (2011) and Dancey and Reidy (2007), correlation coefficients between .30 and .70 reflect a moderate correlation, whereas coefficients greater than .70 indicate a strong correlation. The inter-factor correlation coefficients in Table 11 demonstrate close values to those reported in Rubach and Lazarides (2021) for the original scales.

3.4. Item Analysis

The intent of further item analysis was achieved through an examination of the difference between the lower and upper 27% of the sample by computing the item-total correlation. The relationship between the item scores in the 27% groups and the total scale scores was analyzed in accordance with established literature on the subject (Büyüköztürk, 2011; Tavşancıl, 2014; Tezbaşaran, 2008).

Positive and high correlations indicate that the internal consistency of the scale is maintained and that items can effectively discriminate when the correlation value (r) is greater than or equal to .30, while the significance of the t-test results confirms internal consistency (Büyüköztürk, 2011; Tavşancıl, 2014). T-tests were also used to evaluate mean level differences (Büyüköztürk, 2011; Tezbaşaran, 2008). The results of the correlations and t-tests, which demonstrate the relationship between the item-total correlation values and the lower and upper 27% groups, are presented in Table 12. The item-total correlation values range between $.62 \geq r \geq .79$ and the mean scores for the lower 27% (N=96) and upper 27% (N=96) groups were found to be statistically significant for each item according to the results of the independent t-test (p<.001). These results indicate that the scales are reliable and discriminate effectively.

Table 12. *Item Analysis Results.*

Factors	Second-order Factors	Item	Item Total Correlation (r)	Lower 27% -Upper 27% T-Test
Factor 1	F 1.1	1	.66	14.96*
		2	.72	18.39*
	F 1.2	3	.70	19.63*
		4	.69	18.16*
		5	.73	19.26*
Factor 2	-	6	.72	18.23*
		7	.73	16.82*
		8	.78	22.23*
		9	.74	19.23*
		10	.76	21.94*
Factor3	-	11	.73	17.85*
		12	.74	18.29*
		13	.78	20.27*
Factor 4	-	14	.67	15.03*
		15	.71	18.77*
		16	.71	21.68*
		17	.75	18.69*
Factor 5	F 5.1	18	.73	17.72*
		19	.79	21.50*
		20	.77	20.55*
		21	.77	20.99*

	F 5.2	22	.69	16.10*
		23	.72	17.07*
		24	.62	14.06*
Factor 6	F 6.1	25	.73	18.66*
		26	.76	18.74*
		27	.73	18.40*
	F 6.2	28	.67	16.26*
		29	.69	15.54*

Note. **p*<.001, Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

3.5. Measurement Invariance

The following step involved evaluating the invariance of the instrument across gender groups (as presented in Table 13). Results of the configural invariance analysis in Table 13 indicate that the adapted instrument maintained a consistent structure across gender groups. Additionally, the factor loadings were found to be equivalent across groups, which supports the metric invariance of the items. The scalar invariance of the instrument was determined by evaluating the equivalence of the values of the subjects in regards to the implicit structure and the observed values (Başusta & Gelbal, 2015). Based on changes in the values of CFI, RMSEA/SRMR, it was concluded that the Turkish version of the instrument demonstrated scalar invariance across gender, as reflected by the invariance of the structure, factor loadings, and item intercepts. To determine the significance of these changes, ΔCFI, ΔRMSEA and ΔSRMR values were compared to established thresholds. We considered values of ΔCFI ≤ -.010 and ΔRMSEA ≤ 0.015, or ΔSRMR ≤ 0.030 for step 1 and values of ΔCFI ≤ - 0.010 and ΔRMSEA ≤ 0.015 or ΔSRMR ≤ .010 for step 2 to indicate insignificant changes in the more restrictive model (Chen, 2007).

Table 13. *Indices analyzing measurement invariance of the final factor model.*

	x^2	*df*	CFI	TLI	RMEAS	SRMR
Configural invariance	1484.565	702	.925	.913	.079	.050
Metric invariance	1533.747	725	.922	.913	.079	.061
Scalar invariance	1571.534	748	.921	.914	.079	.063

The multi-group model was established with the objective of determining scalar invariance, which involves the assessment of equivalence in factor structure, factor loadings, and item intercepts. In light of the absence of a specific hypothesis or need to test strict invariance, no such assessment was conducted (Scherer et al., 2017). The results of gender differences for each factor are presented in Table 14.

Based on the probability values, difference for only one competence dimension was determined between male and female teachers in their basic ICT competence beliefs – Male teachers reported higher competence beliefs for digital content creation compared to female teachers.

Table 14. *Gender Differences.*

	Male (n = 108)			Female (n = 248)			*t*	*df*	*p*	*d*	95% CI
	M	*SD*	95% CI	M	*SD*	95% CI					
Factor 1	4.15	.77	[4.01; 4.29]	4.00	.79	[3.90; 4.10]	-1.62	354	.11	-.19	[-.418; .035]
Factor 2	4.16	.79	[4.02; 4.31]	4.04	.87	[3.93; 4.16]	-1.22	354	.22	-.14	[-.368; .084]
Factor 3	3.61	1.15	[3.42; 3.84]	3.32	1.18	[3.16; 3.46]	-2.13	354	.03	-.25	[-.474; -.021]
Factor 4	3.95	.87	[3.80; 4.12]	4.01	.91	[3.89; 4.12]	0.53	354	.60	.07	[-.159; .293]
Factor 5	3.64	.86	[3.49; 3.82]	3.50	.86	[3.35; 3.61]	-1.36	354	.18	-.16	[-.389; .063]
Factor 6	3.65	.97	[3.46; 3.83]	3.44	1.00	[3.30; 3.58]	-1.83	354	.07	-.21	[-.438; .015]

Factor 1 = Information and data literacy, Factor 2 = Communication and collaboration, Factor 3 = Digital content creation, Factor 4 = Safety and security, Factor 5 = Problem-solving, Factor 6 = Analyzing and reflecting.

4. DISCUSSION and CONCLUSION

This study aimed to validate the Turkish version of the "Teachers' Basic ICT Competence Beliefs" instrument developed by Rubach and Lazarides (2021). The underlying structure proposed by Rubach and Lazarides (2021) was adopted as the theoretical framework for this study. Confirmatory factor analysis (CFA) was employed to determine the validity of the Turkish version of the instrument and replicate the six-factor structure, including the second-order structure of three factors.

The subsequent confirmatory analysis revealed that the 29-item Turkish version of the "Teachers' Basic ICT Competence Beliefs" instrument demonstrated acceptable agreement with the original model, as evidenced by high reliability indices. The validity and reliability values of the Turkish scale were comparable to those of the original scale (Rubach & Lazarides, 2021), implying intercultural compatibility for future research. The items of the scale were deemed reliable and distinct for both upper and lower groups, thus suggesting potential benefits for the professional development of in-service and pre-service teachers according to international standards.

Three items from the original scale were removed in the Turkish version as they were found to be inconsistent with the data collected. This discrepancy might be attributed to individuals expressing themselves differently due to intercultural language differences (Ay et al., 2015). Furthermore, variations in factors such as digital technology literacy, access to technology, and usage habits may have contributed to disparities in responses compared to the German sample as described by Koehler & Mishra (2005) and Tondeur, Valcke, & Van Braak (2008). Additionally, personal factors such as attitudes, characteristics, and experiences regarding the utilization of digital technology could also play a role in shaping an individual's ICT (Buabeng-Andoh, 2012). Future research should aim to further understand the psychological processes and similarities and differences in competence beliefs across different cultures, such as Germany and Turkey.

Gender is a crucial individual characteristic that may impact ICT competence beliefs. Thus, it is important to first estimate the invariance of the measurement instrument across gender groups. The results of this study indicated that scalar measurement invariance was approved across gender groups, consistent with the findings of Rubach and Lazarides (2019). In the subsequent analysis, mean-level differences in ICT competence beliefs between male and female teachers were investigated. The results revealed a single difference, with male teachers exhibiting higher competence beliefs in the digital content creation dimension compared to female teachers. Although this difference was observed, it was of a small effect size, which is typical in studies examining gender differences. In the German context, Rubach and Lazarides (2019) found no significant gender differences for the dimensions of information and data literacy, as well as communication and collaboration, but for the dimensions of digital content creation, security, problem solving, and analysis and reflection, male teachers consistently demonstrated higher competence beliefs. These results highlight the potential for intercultural differences in teachers' ICT competence beliefs according to gender. A meta-analysis study by Cai et al., (2017) found that men exhibited more positive attitudes and self-efficacy towards technology use compared to women. It is suggested that future research should further explore the psychological processes and similarities and differences in competence beliefs across different cultures and teacher groups, particularly in the context of the successful use of ICT in education.

The present study has some limitations that need to be considered. Firstly, the sample of teachers was drawn from a single city (Bursa) in Turkey, limiting the generalizability of the results. Furthermore, previous studies that aimed to develop and/or validate instruments in Turkey have mostly focused on pre-service teachers and focused on the level of ICT use, while this study focuses on in-service teachers (Anagün et al., 2016; Gökçearslan et al., 2019; Kutluca et al., 2010; Türel et al., 2017). In the adaptation of the Technological Pedagogical Content Knowledge-Application (TPACKPratik) model for the Turkish culture, items measuring general ICT use were utilized (Ay, Karadağ & Acat, 2015). Future work would benefit from measuring ICT competence beliefs of both in-service and pre-service teachers in line with international criteria.

Secondly, reliability of the instrument was estimated using the Cronbach Alpha Coefficient, the McDonald's Omega, and the Composite Reliability (CR) coefficient. While these coefficients point to acceptable levels of reliability, recent discussions have highlighted the higher value obtained using HTMT2 instead of CR (Roemer et al., 2021). Therefore, future studies might consider the calculation of HTMT2 to increase the robustness of the findings. Additionally, it may be recommended to examine item reliability and assess for different values in multicollinearity in similar studies.

However, despite these limitations, the present study holds significant value in that it has established the validity of the ICT competence beliefs scale as a tool to capture the basic ICT competence beliefs of teachers in Turkey.The present study has found that the ICT competence beliefs scale is a valid instrument to measure teachers' basic ICT competence beliefs in the Turkish context. This result highlights the significance of basic ICT competence beliefs in the utilization of technology in the classroom, as highlighted by various studies (Guggemos & Seufert, 2021; Hatlevik & Hatlevik, 2018; Quast, Rubach, & Lazarides, 2021). The instrument can be used in future research in Turkey to examine the relationship between basic ICT competence beliefs and the actual use of technology by teachers in their professional setting. Additionally, The instrument can be utilized in the realm of teacher education in Turkey to assess existing initiatives aimed at preparing student teachers for the integration of information and communication technology (ICT) in their instructional practices. This will enable the determination of teacher training needs related to ICT, based on international standards. It is necessary to accurately determine educational needs for the enhancement of teachers' beliefs regarding ICT competence, which plays a crucial role in the successful integration of technology in the classroom (Mishra & Koehler, 2006; Yurdakul, Odabasi, Kilicer, Coklar, Birinci, & Kurt, 2012). Consequently, the utilization of this instrument for the needs assessment of current educational programs can aid in the development of effective and efficient programs aimed at enhancing technology integration in the classroom.

Our replication of the six-factor solution of the instrument substantiated its utility in evaluating teachers' fundamental beliefs regarding information and communication technology (ICT) competence in Turkey. Consequently, the adaptation of the instrument into Turkish language has been validated and demonstrated reliability.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s). **Ethics Committee Number**: Yildiz Technical University, Social and Human Sciences Research Ethics Committee, 2021/0l.

Authorship Contribution Statement

Pinar Korukluoglu: Conceptualization, Methodology, Validation, Formal analysis, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing. **Bulent Alci:** Conceptualization, Methodology, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing. **Charlott Rubach:** Conceptualization, Methodology, Validation, Formal analysis, Data Curation, Writing - Review & Editing.

Orcid

Pinar Korukluoglu https://orcid.org/0000-0003-2077-6060
Bulent Alci https://orcid.org/0000-0002-4720-3855
Charlott Rubach https://orcid.org/0000-0003-0451-6429

REFERENCES

Anderson, J., Van Weert, T. (Eds.), & Duchâteau, C. (2002). Information and communication technology in education: A curriculum for schools and program of teacher development. UNESCO. https://pure.unamur.be/ws/portalfiles/portal/256022/129538f.pdf

Anagün, Ş.S., Atalay, N., Kılıç, Z., & Yaşar, S. (2016). Öğretmen adaylarına yönelik 21. yüzyıl becerileri yeterlilik algıları ölçeğinin geliştirilmesi: Geçerlik ve güvenirlik çalışması [The development of a 21st century skills and competences scale directed at teaching candidates: Validity and reliability study]. *Pamukkale University Journal of Education*, *40*(40), 160-175. http://dx.doi.org/10.9779/PUJE768

Ananiadou, K., & Claro, M. (2009). *21st Century Skills and Competences for New Millennium Learners in OECD Countries* (No. 41). OECD Education Working Papers, No. 41.OECD Publishing. http://dx.doi.org/10.1787/218525261154

Angeli, C., & Valanides, N. (2009). Epistemological and methodological issues for the conceptualization, development, and assessment of ICT–TPCK: Advances in technological pedagogical content knowledge (TPCK). *Computers & Education*, *52*(1), 154-168. https://doi.org/10.1016/j.compedu.2008.07.006

Asri, A.A.S.M.A.N., & Darma, G.S. (2020). Revealing the digital leadership spurs in 4.0 industrial revolution. *International Journal of Business, Economics & Management, 3*(1), 93-100. https://doi.org/10.31295/ijbem.v3n1.135.

Ay, Y., Karadağ, E., & Acat, M.B. (2015). The Technological Pedagogical Content Knowledge-practical (TPACK-Practical) model: Examination of its validity in the Turkish culture via structural equation modeling. *Computers & Education*, *88*, 97-108. https://doi.org/10.1016/j.compedu.2015.04.017

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral changes. *Psychol. Rev. 84*, 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bacon, D.R., Sauer, P.L., & Young, M. (1995). Composite reliability in structural equations modeling. *Educational and psychological measurement, 55*(3), 394-406.

Baş, M., & Balaman, F. (2021). Yenilikçi iş davranışı ölçeği'nin Türkçeye uyarlanması: Geçerlik-güvenirlik çalışması [Adaptation of ınnovative work behavior scale to Turkish: A Validity-reliability study]. *Mersin University Journal of the Faculty of Education*, *17*(3).

Başusta, N.B., & Gelbal, S. (2015). Gruplararası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği [Examination of measurement ınvariance at groups’ comparisons: A Study on PISA student questionnaire]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *30*(4), 80-90.

Bayazıt, A., & Seferoğlu, S.S. (2009). Türkiye'deki teknoloji politikalarında eğitimin yeri ve öğretmen yetiştirme politikaları [The place of education in technology policies in Turkey and teacher training policies]. TBD 26. National Informatics Congress, 12. Education Congress in the Light of Information Technologies (BTIE'2009), Proceedings Book, 7-11. Türkiye Bilişim Derneği. http://yunus.hacettepe.edu.tr/~sadi/yayin/BTIE09_Bayazit-Seferoglu_TeknolojiPolitika.pdf

Buabeng-Andoh, C. (2012). Factors influencing teachers adoption and integration of information and communication technology into teaching: A review of the literature. *International Journal of Education and Development using ICT*, *8*(1), 136-155. https://www.learntechlib.org/p/188018/

Büyüköztürk, Ş. (2011). Sosyal bilimler için veri analizi el kitabı istatistik, araştırma deseni, SPSS uygulamaları ve yorum [Data analysis handbook for social sciences statistics, research design, SPSS applications and interpretation] (13th edition). Pegem Akademi.

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education*, *105*, 1-13. https://doi.org/10.1016/j.compedu.2016.11.003

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://link.springer.com/article/10.1007/bf02310555

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2010*). Çok değişkenli istatistik SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi Yayınları.

Dancey, C.P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

Daoud, J.I. (2017, December). Multicollinearity and regression analysis. In *Journal of Physics: Conference Series* (Vol. 949, No. 1, p. 012009). IOP Publishing.

Dündar, S., Ekşi. H, & Yıldız, A. (2008). Aksiyonda değerler ölçeği dilsel eşdeğerlik geçerlik ve güvenirlik çalışması [Modified Values in Action Questionnaire (VIA)]. *Değerler Eğitimi Dergisi*, *6*(15), 89-110.

Eğitim Bilişim Ağı (EBA). (2020). Education Information Network. http://www.eba.gov.tr/

Eickelmann, B., & Vennemann, M. (2017). Teachers 'attitudes and beliefs regarding ICT in teaching and learning in European countries. *European Educational Research Journal*, *16*(6), 733-761. https://doi.org/10.1177/1474904117725899

European Commission (EACEA/Eurydice). (2019). *Digital education at school in Europe. Eurydice Report.* Luxembourg: Publications Office of the European Union. https://eacea.ec.europa.eu/national-policies/eurydice/sites/default/files/en_digital_education_n.pdf

Ferrari, A. (2012). Digital competence in practice: An analysis of frameworks. *Sevilla: JRC IPTS.* http://doi.org/10.2791/82116

Ferrari, A. (2013). *DigComp: A framework for developing and understanding digital competence in Europe. EUR, scientific and technical research series* (Vol. 26035). Luxembourg Publications Office. http://digcomp.org.pl/wp-content/uploads/2016/07/DIGCOMP-1.0-2013.pdf

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th Ed.). Sage.

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age: The IEA International Computer and Information Literacy Study international report* (p. 308). Springer Nature. https://doi.org/10.1007/978-3-319-14222-7

Galanouli, D., Murphy, C., & Gardner, J. (2004). Teachers' perceptions of the effectiveness of ICT-competence training. *Computers & Education*, *43*(1-2), 63-79. https://doi.org/10.1016/j.compedu.2003.12.005

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Allyn & Bacon.

Gerick, J., Eickelmann, B., & Bos, W. (2017). School-level predictors for the use of ICT in schools and students' CIL in international comparison. *Large-scale Assessments in Education*, *5*(1), 1-13. https://doi.org/10.1186/s40536-017-0037-7

Gil-Flores, J., Rodríguez-Santero, J., & Torres-Gordillo, J.J. (2017). Factors that explain the use of ICT in secondary-education classrooms: The role of teacher characteristics and school infrastructure. *Computers in Human Behavior*, *68(1)*, 441-449. https://doi.org/10.1016/j.chb.2016.11.057

Gökçearslan, Ş., Karademir Coşkun, T., & Şahin, S. (2019). Öğretmen adayı bilgi ve iletişim teknolojisi yeterlikleri ölçeğinin Türkçe'ye uyarlanması [Adaptation of information and communication technology competency scale to Turkish for pre-service teachers]. *Kastamonu Education Journal, 27*(4), 1435-1444. https://doi.org/10.24106/kefdergi.2828

Guggemos, J., & Seufert, S. (2021). Teaching with and teaching about technology–Evidence for professional development of in-service teachers. *Computers in Human Behavior*, *115*, 106613. https://doi.org/10.1016/j.chb.2020.106613

Hatlevik, O.E. (2017). Examining the relationship between teachers' self-efficacy, their digital competence, strategies to evaluate information, and use of ICT at school. *Scandinavian Journal of Educational Research*, *61*(5), 555-567. https://doi.org/10.1080/00313831.2016.1172501

Hatlevik, I.K., & Hatlevik, O.E. (2018). Examining the relationship between teachers' ICT self-efficacy for educational purposes, collegial collaboration, lack of facilitation and the use of ICT in teaching practice. *Frontiers in psychology*, *9*, 935. https://doi.org/10.3389/fpsyg.2018.00935

Hatlevik, O.E., Throndsen, I., Loi, M., & Gudmundsdottir, G.B. (2018). Students' ICT self-efficacy and computer and information literacy: Determinants and relationships. *Computers & Education*, *118*, 107-119. https://doi.org/10.1016/j.compedu.2017.11.011

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

Hutcheson, G.D., & Sofroniou, N. (1999). *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models*. SAGE.

Ilomäki, L., Paavola, S., Lakkala, M. *et al.* (2016). (Digital competence – an emergent boundary concept for policy and educational research. *Educ Inf Technol.*, 21(1), 655–679 https://doi.org/10.1007/s10639-014-9346-4

International Journal of Assessment Tools in Education (IJATE). (2014). Guidelines for Translating and Adapting Tests. https://dergipark.org.tr/tr/pub/ijate

International Society for Technology in Education (ISTE). (2008). Essential conditions: Necessary conditions to effectively leverage technology for learning. https://www.iste.org/standards/for-educators

İzmirli, Ö.Ş. (2015). Öğretim Sürecine BİT Entegrasyonunu Etkinlik Kuramı Çerçevesinde Anlama: Bir Durum Çalışması [Understanding ICT integration into instructional processes within the scope of activity system theory: A case study]. *Education and Science*, *40*(180), 307-325. http://dx.doi.org/10.15390/EB.2015.4725

Kapur, R., Byfield, V., Del Frate, F., Higgins, M., & Jagannathan, S. (2018). The digital transformation of education. *Earth Observation Open Science and Innovation*, *15*, 25-41. http://dx.doi.org/10.1007/978-3-319-65633-5

Kılıç, S. (2016). Cronbach's alpha reliability coefficient. *Journal of Mood Disorders*, *6*(1), 47-48. http://dx.doi.org/10.5455/jmood.20160307122823

Kılıç, C., & Alcı, B. (2022). Sosyal duygusal öğrenme ölçme aracının Türkçe 'ye uyarlanması: geçerlilik ve güvenilirlik çalışması [Validity and reliability study of Turkish form of social emotional learning questionnaire]. *e-Uluslararası Eğitim Araştırmaları Dergisi, 13* (1), 38-50. https://doi.org/10.19160/e-ijer.975137

Kim, H.Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52-54. https://doi.org/10.5395/rde.2013.38.1.52

Kocaman Karoğlu, A., Bal, K., & Çimşir, E. (2020). Toplum 5.0 sürecinde Türkiye'de eğitimde dijital dönüşüm [Digital transformation of education in Turkey in society 5.0]. *Journal of University Research, 3*(3), 147-158. https://doi.org/10.26701/uad.815428

Koehler, M.J., & Mishra, P. (2005). What happens when teachers design educational technology? The development of technological pedagogical content knowledge. *Journal of Educational Computing Research*, *32*(2), 131-152. https://doi.org/10.2190/0EW7-01WB-BKHL-QDYV

Koh, J.H.L., Chai, C.S., & Tsai, C.C. (2013). Examining practicing teachers' perceptions of technological pedagogical content knowledge (TPACK) pathways: A structural equation modeling approach. *Instructional Science, 41*(4), 793-809. https://doi.org/10.1007/s11251-012-9249-y

Kultusministerkonferenz (2016). *Bildung in der digitalen Welt: Strategie der Kultusministerkonferenz* [Education in digital environment: Strategy of the Kultusministerkonferenz]. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2018/Strategie_Bildung_in_der_digitalen_Welt_idF._vom_07.12.2017.pdf

Kutluca, T., Arslan, S., & Özpinar, İ. (2010). Developing a scale to measure information and communication technology utilization levels. *Journal of Turkish Science Education*, *7*(4), 37-45. http://tused.org/index.php/tused/article/view/535

Krumsvik, R.J. (2014). Teacher educators' digital competence. *Scandinavian Journal of Educational Research, 58*(3), 269–280. https://doi.org/10.1080/00313831.2012.726273

McDonald, R.P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Ministary of National Education (MoNE) [Milli Eğitim Bakanlığı]. (2006). Temel eğitime destek projesi "öğretmen eğitimi bileşeni" öğretmenlik mesleği genel yeterlikleri [Basic education support project "teacher training component" general competencies of the teaching profession]. *Tebliğler Dergisi*, *2590*, 1491-1540. http://oygm.meb.gov.tr/meb_iys_dosyalar/2017_12/13161921_YYretmenlik_MesleYi_Genel__YETERLYKLERi_onaylanan.pdf

Ministary of National Education (MoNE) [Milli Eğitim Bakanlığı]. (2017). Öğretmenlik Mesleği Genel Yeterlikleri [General Competencies of Teaching Profession]. *Tebliğler Dergisi.* https://oygm.meb.gov.tr/meb_iys_dosyalar/2017_12/11115355_YYRETMENLYK_MESLEYY_GENEL_YETERLYKLERY.pdf

Mishra, P., & Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for integrating technology in teachers' knowledge. *Teachers College Record, 108*(6), 1017-1054. https://www.learntechlib.org/p/99246/.

Muenks, K., Wigfield, A., & Eccles, J.S. (2018). I can do this! The development and calibration of children's expectations for success and competence beliefs. *Developmental Review*, *48*, 24-39. https://doi.org/10.1016/j.dr.2018.04.001

Muth´en, L.K., & Muth´en, B.O. (1998-2016). *Mplus 8.1* [computer software]. Los Angeles, CA: Muth´en & Muth´en.

Organisation for Economic Cooperation and Development (OECD). (2018). *The future of education and skills: Education 2030.* OECD Education Working Papers. http://www.oecd.org/education/2030/oecd-education-2030-position-paper.pdf

Öğretmen Bilişim Ağı (ÖBA). (2022). Teacher Information Network. https://www.oba.gov.tr/

Palvia, S., Aeron, P., Gupta, P., Mahapatra, D., Parida, R., Rosner, R., & Sindhi, S. (2018). Online education: Worldwide status, challenges, trends, and implications. *Journal of Global Information Technology Management, 21*(4). 233-241. https://doi.org/10.1080/1097198X.2018.1542262

Parlak, B. (2017). Dijital çağda eğitim: olanaklar ve uygulamalar üzerine bir analiz [Education in digital age: An analysis on opportunities and applications] *Suleyman Demirel University the Journal of Faculty of Economics and Administrative Sciences, Special Issue on Kayfor, 15*, 1741-1759.

Patton, M.Q. (2018). *Nitel araştırma ve değerlendirme yöntemleri* [Qualitative research and evaluation methods]. (Demir,S.B. & Bütün, M., trans.). Pegem Akademi

Quast, J., Rubach, C., & Lazarides, R. (2021). Lehrkräfteeinschätzungen zu Unterrichtsqualität mit digitalen Medien: Zusammenhänge zur wahrgenommenen technischen Schulausstattung, Medienunterstützung,digitalen Kompetenzselbsteinschätzungen und Wertüberzeugungen [Teachers' perceptions of quality teaching with ICT: Associations with perceived school ICT equipment, ICT support, ICT competence beliefs and value beliefs]. *Zeitschrift für Bildungsforschung*. https://doi.org/10.1007/s35834-021-00313-7

Redecker, C., & Punie, Y. (2017). Digital Competence of Educators. JRC Science for Policy Report. https://core.ac.uk/download/pdf/132627227.pdf

Roemer, E., Schuberth, F., & Henseler, J. (2021). HTMT2–an improved criterion for assessing discriminant validity in structural equation modeling. *Industrial management & data systems*.

Rubach, C., & Lazarides, R. (2019). Eine Skala zur Selbsteinsch¨atzung digitaler Kompetenzen bei Lehramtsstudierenden: Entwicklung eines Instrumentes und die Validierung durch Konstrukte zur Mediennutzung und Werteüberzeugungen zur Nutzung digitaler Medien im Unterricht [A scale for self-percieved digital competence for student teachers: Development and validation with constructs like ICT, ICT values]. *Zeitschrift Für Bildungsforschung, 2*(78), 4. https://doi.org/10.1007/s35834-019-00248-0

Rubach, C., & Lazarides, R. (2021). Addressing 21st-century digital skills in schools–Development and validation of an instrument to measure teachers' basic ICT competence beliefs. *Computers in Human Behavior*, *118*, 106636. https://doi.org/10.1016/j.chb.2020.106636

Seufert, S., Guggemos, J., & Sailer, M. (2021). Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, *115*, 106552. https://doi.org/10.1016/j.chb.2020.106552

Scherer, R., Tondeur, J., & Siddiq, F. (2017). On the quest for validity: Testing the factor structure and measurement invariance of the technology-dimensions in the Technological, Pedagogical, and Content Knowledge (TPACK) model. *Computers & Education, 112*, 1–17. https://doi.org/10.1016/j.compedu.2017.04.012

Seçer, İ. (2018*). Psikolojik test geliştirme ve uyarlama süreci SPSS ve LISRELL uygulamaları* [Psychological test development and adaptation process SPSS and LISRELL applications]. (2th edition). Anı Yayıncılık.

Şad, S.N., & Nalçacı, Ö.İ. (2015). Öğretmen Adaylarının Eğitimde Bilgi ve İletişim Teknolojilerini Kullanmaya İlişkin Yeterlilik Algıları [Prospective Teachers' Perceived Competencies about Integrating Information and Communication Technologies into

Education]. *Mersin University Journal of the Faculty of Education*, *11*(1). http://abakus.inonu.edu.tr/xmlui/handle/123456789/17255

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics*, 6th Ed. Pearson.

Tavşancıl, E. (2014). *Tutumların ölçülmesi ve SPSS ile veri analizi*. Ankara: Nobel Yayın Dağıtım

Tekindal, S. (2015). *Duyuşsal özelliklerin ölçülmesi için araç oluşturma* [Creating a tool for measuring affective traits] (3rd ed.). Pegem A Akademi.

Tezbaşaran, A. (2008). Likert tipi ölçek hazırlama kılavuzu. https://www.academia.edu/1288035/Likert_Tipi_Ölçek_Hazırlama_Kılavuzu

Tondeur, J., Aesaert, K., Prestridge, S., & Consuegra, E. (2018). A multilevel analysis of what matters in the training of pre-service teacher's ICT competencies. *Computers & Education*, *122*, 32-42. https://doi.org/10.1016/j.compedu.2018.03.002

Tondeur, J., Van Braak, J., & Valcke, M. (2007). Curricula and the use of ICT in education: Two worlds apart? *British Journal of educational technology*, *38*(6), 962-976. https://doi.org/10.1111/j.1467-8535.2006.00680.x

Tondeur, J., Valcke, M., & Van Braak, J. (2008). A multidimensional approach to determinants of computer use in primary education: Teacher and school characteristics. *Journal of computer assisted learning*, *24*(6), 494-506. https://doi.org/10.1111/j.1365-2729.2008.00285.x

Trochim, W.M. & Donnelly, J.P. (2006). The research methods knowledge base. 3rd Edition, Atomic Dog, Cincinnati, OH.

Turkish Academy of Sciences (TÜBA) (Türkiye Bilimler Akademisi) (2002). *Bilimsel araştırmada etik ve sorunları* [Ethics and problems in scientific research]. TÜBA. http://www.tuba.gov.tr/tr/yayinlar/suresiz-yayinlar/raporlar/tuba-bilimsel-arastirmada-etik-ve-sorunlari

Türel, Y.K., Özdemir, T.Y., & Varol, F. (2017). Öğretmenlerin bilgi ve iletişim teknolojileri becerileri ölçeği: Güvenirlik ve geçerlik [Teachers' ICT skills scale (TICTS): Reliability and validity]. *Çukurova University Faculty of Educational Journal, 46*(2), 503-516. http://doi.org/10.14812/cuefd.299864

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2011). ICT Competency Framework for Teachers. https://unesdoc.unesco.org/ark:/48223/pf0000213475.

Vanderlinde, R., & Van Braak, J. (2010). The e-capacity of primary schools: Development of a conceptual model and scale construction from a school improvement perspective. *Computers & Education, 55*(2), 541-553. https://doi.org/10.1016/j.compedu.2010.02.016

Voogt, J., & Roblin, N.P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299-321. https://doi.org/10.1080/00220272.2012.668938

Wang, H.Y., Sigerson, L., & Cheng, C. (2019). Digital nativity and information technology addiction: Age cohort versus individual difference approaches. *Computers in Human Behavior*, *90*, 1-9. https://doi.org/10.1016/j.chb.2018.08.031

Wicherts, J.M. (2007). Group differences in intelligence test performance. Unpublished dissertation, University of Amsterdam.

Yurdakul, K.I., Dönmez, O., Altınok, A., & Odabaşı, H.F. (2013). Dijital ebeveynlik ve değişen roller [Digital parenting and changing roles]. *Gaziantep University Journal of Social Sciences*, *12*(4), 883-896.

Yurdakul, I.K., Odabasi, H.F., Kilicer, K., Coklar, A.N., Birinci, G., & Kurt, A.A. (2012). The development, validity and reliability of TPACK-deep: A technological pedagogical content knowledge scale. *Computers & Education*, *58*(3), 964-977. https://doi.org/10.1016/j.compedu.2011.10.012

Zhong, L. (2017). Indicators of digital leadership in the context of K-12 education. *Journal of Educational Technology Development and Exchange (JETDE), 10*(1), 27- 40.

APPENDIX

Translation of All Items for Each Basic ICT Competence Dimension.

ITEM WORDING IN GERMAN (original version)	ITEM WORDING IN ENGLISH (original version)	ITEM WORDING IN TURKISH
	Factor 1: Information and data literacy	**Boyut 1: Bilgi Veri Okuryazarlığı**
Ich kann auf Grundlage meiner Suchinteressen relevante Quellen in digitalen Umgebungen identifizieren und nutzen.	I can identify and use appropriate sources in digital environments based on my information needs.	Dijital ortamdaki kaynakları bilgi ihtiyacıma göre belirleyip uygun bir şekilde kullanabilirim
Ich kann Suchstrategien im digitalen Raum nutzen.	I can use my search strategies in digital environments.	Araştırma stratejilerimi dijital ortamda kullanabilirim
Ich kann Informationen, Informationsquellen und Daten im digitalen Raum kritisch bewerten.	I am critical about information, sources and data in digital environments.	*Excluded from scale*
Ich kann digital Informationen und Daten sicher speichern.	I can store digital information and data securely.	Dijital bilgi ve verileri güvenli bir şekilde depolayabilirim.
Ich kann Informationen, die ich gespeichert habe, wiederfinden.	I can retrieve the information that I have stored.	Depoladığım bilgileri geri getirebilirim.
Ich kann Informationen, die ich gespeichert habe, von verschiedenen Orten abrufen.	I can retrieve information that I have stored from different environments.	Farklı ortamlardan depoladığım bilgileri geri getirebilirim.
	Factor 2: Communication and collaboration	**Boyut 2: İletişim ve İşbirliği**
Ich kann mit Hilfe verschiedener digitaler Medien kommunizieren.	I can communicate using different digital media.	Farklı dijital medyaları kullanarak iletişim kurabilirim
Ich kann Informationen und Dateien aus dem digitalen Raum zitieren.	I can cite information and files from digital environments.	Dijital ortamlardan bilgi ve dosya alıntılayabilirim
Ich kann digitale Medien nutzen, um gemeinsam mit anderen Dateien und Dokumente zu bearbeiten.	I can edit files and documents collaboratively with others using digital media	Dijital ortamları kullanarak, dosyaları ve belgeleri başkalarıyla birlikte düzenleyebilirim
Ich kann Verhaltensregeln bei digitalen Interaktionen und Kooperationen anwenden.	I can apply behavioral rules in digital interactions and collaborations.	Dijital etkileşim ve işbirliği konularında davranışsal kuralları uygulayabilirim
Ich kann mit Hilfe digitaler Medien aktiv an der Gesellschaft teilhaben.	I can actively participate in society using digital media.	Dijital medyayı kullanarak, topluma aktif bir şekilde katılabilirim.
Ich kann meine Medienerfahrungen in Interaktion mit anderen weitergeben.	I can share my experiences with digital media in interactions with others	*Excluded from scale*
	Factor 3: Digital content creation	**Boyut 3: Dijital İçerik Oluşturma**
Ich kann mir bekannte Apps und Programme bedarfsgerecht anwenden.	I can use familiar apps and programs according to my needs.	*Excluded from scale*
Ich kann eigene digitale Produkte in verschiedenen Formaten gestalten.	I can design my digital products in various formats.	Dijital ürünlerimi çeşitli formatlarda tasarlayabilirim.
Ich kann digitale Inhalte in verschiedenen Formaten bearbeiten und zusammenführen	I can edit and merge digital content in different formats	Dijital içerikleri, farklı formatlarda düzenleyebilir ve birleştirebilirim

Ich kann digitale Inhalte in verschiedenen Formaten pr¨asentieren.	I can present digital content in different formats.	Dijital içeriği farklı formatlarda sunabilirim
	Factor 4: Safety and security	**Boyut 4: Emniyet ve Güvenlik**
Ich kenne die Gefahren und Risiken in digitalen Umgebungen und berücksichtige diese.	I know about the dangers and risks in digital environments and consider them.	Dijital ortamlardaki tehlike ve riskleri bilir ve bunları dikkate alırım.
Ich kann meine Privatsph¨are in digitalen Umgebungen durch geeignete Maßnahmen schützen.	I can protect my privacy in digital environments through appropriate measures.	Dijital ortamlarda gizliliğimi gerekli önlemler aracılığıyla koruyabilirim
Ich kann meine Sicherheitseinstellungen regelm¨aßig aktualisieren.	I can regularly update my security settings.	Güvenlik ayarlarımı düzenli olarak güncelleyebilirim.
Ich kann digitale Technologien gesundheits- und umweltbewusst nutzen.	I can use digital technologies in a healthy and environmentally sound way.	Dijital teknolojileri sağlıklı ve çevreye duyarlı bir şekilde kullanabilirim
	Factor 5: Problem Solving	**Boyut 5: Problem Çözme**
Ich kann digitale Werkzeuge, Tools und Plattformen bedarfsgerecht einsetzen	I can use digital tools and platforms according to my needs.	Dijital araç ve platformları ihtiyaçlarım doğrultusunda kullanabilirim
Ich kann digitale Werkzeuge zum pers¨onlichen Gebrauch anpassen	I can adapt digital tools for personal use.	Dijital araçları kişisel kullanımıma göre uyarlayabilirim
Ich kann digitale Lernm¨oglichkeiten und dafür geeignete Tools selbstst¨andig nutzen.	I can independently use digital learning opportunities and appropriate tools	Dijital öğrenme imkanlarını ve uygun araçları bağımsız bir şekilde kullanabilirim.
Ich kann digitale Lernressourcen selbstst¨andig organisieren.	I can organize digital learning resources independently.	Dijital öğrenme kaynaklarını bağımsız bir şekilde düzenleyebilirim
Ich kann L¨osungen für technische Probleme entwickeln.	I can develop solutions for technical problems.	Teknik sorunlara karşı çözüm üretebilirim.
Ich kenne Funktionsweisen und grundlegende Prinzipien des digitalen Raumes.	I know about the functioning and basic principles of digital systems.	Dijital sistemlerin işleyişi ve temel ilkeleri hakkında bilgiye sahibim.
Ich erkenne algorithmische Strukturen bei genutzten Tools.	I identify algorithmic structures in the tools I use.	Kullandığım araçlardaki algoritmik yapıları tanımlarım.
	Factor 6: Analyzing and reflecting	**Boyut 6: İnceleme ve Yansıtma**
Ich kann die Wirkung von Medien im digitalen Raum analysieren.	I can analyze the effect of media in digital environments.	Dijital ortamlarda medyanın etkisini analiz edebilirim
Ich kann eine interessengeleitete Verbreitungen und die Dominanz von Themen im digitalen Raum beurteilen.	I can evaluate interest-driven dissemination and the dominance of topics in digital space.	Dijital alanda ilgi odaklı bilgi yayılmasını ve konu baskınlığını değerlendirebilirim
Ich kann Chancen und Risiken des Mediengebrauchs für meinen eigenen Mediengebrauch reflektieren.	I can reflect on the opportunities and risks of media use for my own media use.	Kişisel medya kullanımım için medya kullanımına dair imkan ve riskleri iyi bir şekilde değerlendirebilirim
Ich kann Vorteile von Gesch¨aftsaktivit¨aten und Services im digitalen Raum analysieren	I can analyze the benefits of business activities and services in digital environments.	Dijital ortamlardaki ticari faaliyetlerin ve hizmetlerin faydalarını analiz edebilirim
Ich kann Risiken von Gesch¨aftsaktivit¨aten und Services im digitalen Raum analysieren.	I can analyze the risks of business activities and services in the digital space.	Dijital ortamlardaki ticari faaliyetlerin ve hizmetlerin risklerini analiz edebilirim

Note: (Original version: Rubach & Lazarides, 2021)

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 56–75

https://doi.org/10.21449/ijate.981367

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

Comparison of Kernel equating methods under NEAT and NEC designs

Seyma Nur Ozsoy[1,*], **Sevilay Kilmen**[2]

[1]Ankara University, PhD Student of Educational Sciences, Ankara, Türkiye
[2]Bolu Abant İzzet Baysal University, Faculty of Education, Department of Educational Sciences, Bolu, Türkiye

ARTICLE HISTORY

Received: Aug. 11, 2021
Accepted: Feb. 02, 2023

Keywords:
Test equating,
Kernel equating,
NEAT design,
NEC design,
Bandwidth.

Abstract: In this study, Kernel test equating methods were compared under NEAT and NEC designs. In NEAT design, Kernel post-stratification and chain equating methods taking into account optimal and large bandwidths were compared. In the NEC design, gender and/or computer/tablet use was considered as a covariate, and Kernel test equating methods were performed by using these covariates and considering bandwidths. The study shows that, in the NEAT design, Kernel chain equating methods exhibit higher error than the post-stratification equating methods do since the lowest error in the NEC design was obtained from the Kernel equating method with large bandwidth through the computer/tablet variable. Kernel test equating results based on the NEC design, which considers gender and computer tablet use variables as a covariate separately, showed lower SEE than that of the NEC pattern, which takes these variables together as covariates. In terms of the bandwidth, when all methods are compared within the pattern used (i.e., NEAT and NEC), it has been seen that generally Kernel test equating with large bandwidth results in fewer errors than the Kernel test equating with optimal bandwidth. When the NEAT and NEC designs are compared generally, the NEAT design has a lower SEE than that of the NEC design.

1. INTRODUCTION

In some testing practices, different test forms are used in different groups to provide test reliability. These tests consisting of different items bring along some equivalence discussions due to varying difficulties. Therefore, the need to equate tests arises in order to prevent injustice in comparing tests.

The concept of test equating has been defined and studied by many researchers for many years and still continues to be among the current research (Kolen & Brennan, 2004; von Davier et al., 2004b; Livingston, 2014). Test equating is accepted as a statistical process used by individuals who are subjected to the same assessment process to make the scores obtained from many forms of this assessment into comparable state (von Davier, 2013; Kolen & Brennan, 2004) since such a process eliminates discussions about which form of test individuals will take because differences between the obtained scores depending on the test form are prevented (Lord, 1980).

*CONTACT: Seyma Nur Ozsoy ✉ 1seymaozsoy@gmail.com Ankara University, phD Student of Educational Sciences, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

Test equating is mainly divided into two categories, namely equating with observed score and true score (Lord, 1980). The observed scores equating is performed with just observed scores and includes equal percentage equating and chain equating approaches (Kolen & Brennan, 1995). On the other hand, in the true score equating, the true score covers the observed score and the standard error. Among the scaling/calibration methods for true score equating, there are approaches such as mean-mean, mean-standard deviation, and Stocking-Lord (Kolen & Brennan, 2004).

Both true score and observed score equating possess limitations. As the true score equating requires assumptions such as large sample size and local independence, in practice using it to equate different test forms can be too hard while in observed score equating using discrete distributions can cause increase in equating errors. To overcome these limitations, the Kernel equating method, a relatively new method, is recommended as an observed score equation method in which score distributions are equated by converting discrete score distributions into continuous distributions by using Gauss Kernel approach instead of the linear approach (von Davier et al., 2004a) because Kernel equating offers more realistic assumptions than the other methods do (Godfrey, 2007). Furthermore, due to the pre-smoothing, Kernel equating gives less standard error compared to other methods, is less dependent on sample size, and can be applied to all designs and equating functions (von Davier et al., 2004b).

On the other hand, test equating generally requires applying an anchor test to different groups that take different tests. This test equating design is called a nonequivalent groups anchor test (NEAT). However, specifically in examinations that are applied several times in a year or term, using the same anchor test sometimes can cause some problems; for example, the use of the same items repeatedly can lead to recall of items for individuals, which can negatively affect discrimination. Recently, as a solution to this problem, there are studies suggesting that test equating can be conducted by using nonequivalent groups with covariates (NEC) design (e.g., Akın Arıkan, 2020; Albano & Wiberg, 2019; Branberg, 2010; Branberg & Wiberg, 2011; Gonzales et al., 2015; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017). For example, Yurtçu (2018) equated scores obtained from different tests by using common item scores, gender, and mathematics self-efficacy scores as covariates. Their results showed that common variables could be used instead of common items to equate test scores obtained from different tests. Akın Arıkan (2020) compared NEAT design and NEC designs using gender and socioeconomic status variables as covariance variables and their study results indicated that NEC design could be taken as a practically viable alternative to the NEAT design in Kernel equating to establish the comparability of the test scores. Notwithstanding the proven utility of the NEC design for obtaining comparable test scores from different groups under the Kernel equating in a limited number of studies, it still remains a question about whether this approach can be used instead of anchor items. Therefore, there is still need for more studies that compare Kernel equating results in NEC design and NEAT design. To this end, the present study focuses on comparing the performance of Kernel test equating methods under NEAT and NEC design.

1.1. NEAT and NEC Design in Test Equating

1.1.1. *Nonequivalent groups with anchor tests (NEAT) design*

In NEAT design, common items in different forms are used to equate test scores obtained from different tests as can be seen in Figure 1. These forms are applied in different groups who do not know the equivalence due to such features as the number of individuals and item order (von Davier et al., 2004b). Anchor test is prepared in accordance with the characteristics of the main test forms. For these common substances to have a similar effect in both forms, the item numbers must also be the same (Kolen, 1988; Kolen & Brennan, 2004). Two test forms are equated by using the anchor test.

Figure 1. *NEAT design.*

	Form A	Anchor test	Form B
Group 1	✓	✓	
Group 2		✓	✓

Post-stratification and chain equating in NEAT design: In this design, information from the anchor items can be provided by two different approaches, namely post-stratification or chain equating approaches. The approach in which anchor test score is used as a conditioning variable (or a covariate) for estimating the score distributions is called the poststratification approach. In this method, the conditional distributions of the X form given anchor test and of the Y form given anchor test are weighted by distribution for anchor test to estimate the score distributions for X form and Y form in a hypothetical target population (T) (von Davier & Chen, 2013). In T denoted as (wP + (1 – w)Q), w is the proportion of T that comes from P, (Braun & Holland, 1982). The second approach, the chain equating approach (von Davier et al., 2006), involves a two-stage process for the transformation of the scores of form X into scores of form Y (von Davier et al., 2004a). In Kernel chain equating, first, the X form is linked to the common items and then the common items are linked to the Y form to ensure equating (Andersson et al., 2013). An important difference between post-stratification equating and chain equating is that in the former there is an explicit target population (T) whereas in the latter T plays no explicit role (von Davier & Chen, 2013). In the present study, Kernelpost-stratification and chain equating methods under NEAT design were used.

1.1.2. *Nonequivalents groups with covariates (NEC) design*

NEAT design may not be used in many test applications for such reasons as test security and recognizing the items which are used in the anchor test of previous test applications (Wiberg, 2015). Branberg and Wiberg (2011) recommended using covariate variables to equate the two different test forms and conducted various studies using the NEC design. In the NEC design, the scores obtained from different tests are equated with the covariate variable/s associated with the test scores (see Figure 2). Covariates are considered similar to the common item scores used in the NEAT design (Wiberg, 2015). The most important feature of covariates is that they are categorical. In many studies where continuous variables are used, the variables are categorical by methods such as cluster analysis. Therefore, in the present study, gender and having a computer/tablet as variables are discussed. The gender variable is important because it is related to the learned roles of women and men in the field of science, and the computer/tablet use variable is important because it allows access to today's information.

Figure 2. *NEC design.*

	Form A		Form B
Group 1	✓	Covariate variable/s	
Group 2			✓

1.2. Kernel Test Equating in NEAT and NEC Design

Kernel equating was first recognized by Livingston (1993) with his test equating study using log-linear smoothing. Kernel equating method is an observed score equation method in which score distributions are equated by converting discrete score distributions into continuous distributions. In these conversions Kernel equating uses Gauss Kernel approach instead of linear approach which is used in the traditional observed score equating (von Davier et al., 2004a). Kernel equating is preferred to traditional test equating methods for at least four reasons: The first is that it has realistic assumptions than other methods (Godfrey, 2007); the second is that due to the pre-smoothing, it gives less standard equating error compared to other methods; the third is that it is less dependent on sample size; and lastly, it can be applied to all designs and equating functions (von Davier et al., 2004b). Kernel equating is carried out in a five-step process (von Davier et al., 2004b), which includes pre-smoothing, estimation of score probabilities, continuization, equating, and calculation of equaling error. Kernel test equating process was explained for both NEAT and NEC designs separately as consistent with the aim of this study.

In the first step of Kernel equating, pre-smoothing is performed in order to reduce complexities in the observed score distributions depending on the sampling. In this step, the data are linked with log linear model (von Davier et al., 2004a). This process is the same in both NEAT and NEC designs. In the second step, score probability is estimated. Score probability estimation varies according to the equating design used as mentioned before. In NEAT design, score probability is estimated by common items, while in NEC design it is estimated by common categorical variable/s. Moreover, when score probabilities are estimated by using anchor test in the NEAT design, two different approaches are used, namely poststratification equating and chain equating (von Davier et al., 2006). In the present study, both approaches were used to see the possible effects of these approaches on the SEEs and to compare them. In the third step, discrete score distributions are made continuous. This process is performed in order to produce two cumulative frequency distributions. Gauss Kernel is commonly used to make the discrete distributions continuous in Kernel equating studies. In addition, in this step, the bandwidth (h parameter) is determined to make the discrete distributions continuous (Gonzales & Wiberg, 2017). The bandwidth can be chosen in two ways as optimal or large bandwidths (von Davier et al., 2006). In the current study, both optimal and large bandwidths were used to see the possible effect of the bandwidth on SEE results. In the fourth step, equating is performed between continuous distributions by using the Kernel equating methods. The Kernel equating function in which an X form is equal to the Y form is as follows (Andersson et al., 2013):

$$\hat{e}_y(x) = G_{hy}^{-1}(F_{hx}(x;\hat{r});\hat{s})$$

$$= G_{hy}^{-1}\left(F_{hx}(x_j)\right)$$

F_{hx} and G_{hy}: Cumulative distribution function
hx and hy: Bandwidths for test x and test y
r and s: Score probabilities for test x and test y

In the last step, equating error is obtained by calculating SEEs in Kernel equating. The SEE obtained by equating the X form to the Y form is calculated using the equation below (Andersson et al., 2013; Gonzales & Wiberg, 2017:

$$SEE_Y(x) = \sqrt{Var\left(\hat{e}_Y(x)\right)}$$

$$SEE_X(y) = \sqrt{Var(\hat{e}_X(y)}$$

1.3. Studies Comparing NEAT and NEC Design

In Kernel test equating studies, it was seen that the NEAT design was commonly used. Over the last decade, NEC design with Kernel equating has been used; for example, Branberg (2010) investigated the use of NEC design in test-equating studies and obtained important findings of the use of covariates in the absence of an anchor test. In another study conducted by Branberg and Wiberg (2011), it was revealed with simulated data that the variables of gender and educational status can reduce the amount of test equating error. Strong evidence was also obtained showing that covariate variable/s can be used to equate different tests. In another study conducted by Gonzales, Barrientos, and Quintana (2015) gender and school type were used as covariates in NEC design. The results presented supportive evidence to previous studies that revealed that covariates can be used in test equating studies. Wiberg and Branberg (2015) compared equated scores obtained from NEC design and equated group design and their study results showed that when common variables are used together with common items, they give fewer errors. Wiberg and von Davier (2017) examined anchor tests using age, gender, and education as covariates. The results obtained in their study indicated that even if the composition of the group taking the exam changes, test results can be controlled. In the study conducted by Albano and Wiberg (2019), in which gender was used as a common variable, it was determined that frequency estimation gives less error in the presence of anchor test and covariate variables. Moreover, recent studies comparing NEAT and NEC designs show that common variables can be used instead of common items. For example, in a test equating study conducted by Yurtçu (2018), gender and mathematics self-efficacy scores were used to equate test scores besides the anchor test and the study results presented evidence that common variables can be used instead of common items. Akın Arıkan (2020) made a comparison of the NEAT design and NEC design using gender and socioeconomic status variables as covariance variables and concluded that in the absence of anchor tests, equating can be made by using covariate variables.

In sum, such studies examined NEC design and compared NEC design with NEAT design to find an alternative to anchor tests in conditions in which NEAT design cannot be used. With an aim to contribute to these studies, in this current study, Kernel equating methods under both NEAT and NEC designs were compared according to their standard errors of equating (SEE). Two booklets numbered 1 and 14 out of 14 different booklets used in the Türkiye sample of the TIMSS 8th grade science test applied in 2019 were used to compare Kernel equating methods under both NEAT and NEC designs. In this present study, gender is considered as a covariate variable for the NEC design. In addition to gender, considering the transition to eTIMSS application in 2019, the use of a computer/tablet use is also considered as a covariate variable.

1.4. The Present Study

In the current study, in NEAT design post-stratification equating and chain equating were compared and in NEC design, gender and computer/tablet use were considered as covariates. On the other hand, the selection of bandwidth was considered as a variable that could affect SEEs. Two bandwidths were used in this study, namely optimal and large bandwidths. Depending on these conditions, ten Kernel test equation SEEs were examined (see Table 1). Consequently, the present study examines the role of test equating methods, bandwidths, and the use of covariate/s on SEEs. Accordingly, three research questions are formulated:

1) Which Kernel test equating method gives less SEE when equating scores are obtained from different TIMSS booklets under NEAT design?
2) Which covariate gives less SEE when equating scores are obtained from different TIMSS booklets under NEC design?
3) How does the selection of bandwidth (optimal and large bandwidth) affect SEE when equating scores are obtained from different TIMSS booklets in both NEAT and NEC design?

As mentioned earlier, the current study focuses on comparing ten Kernel test equating conditions under NEAT and NEC designs according to their SEEs. Examining the SEEs between different Kernel equating methods under different test equating designs is crucial for at least three reasons. First, it has been known that different Kernel test equating results give different SEEs. To compare these results and create some advice about which test equating methods are more proper and in which situation, these test equating methods should be examined in various test conditions. Therefore, there is need to conduct further studies addressing test equating method comparisons in various test conditions. The current study therefore compares test equating methods by focusing on Kernel test equating, which is used under conditions that can be met in practice.

Second, although there is a number of studies that compare Kernel equating methods, except for limited research (e.g., Choi, 2009; Liang & von Davier, 2014), there is lack of research examining the performance of Kernel equating regarding the choice of bandwidth and how choices on bandwidth affect equating results in terms of SEE. Relevant literature shows that the bandwidth parameter determines the smoothness of the continuized score distributions and has a large effect on the Kernel density estimate. Relevant research results also show that there is a need to investigate how the bandwidths affect the equating results more rigorously and also to identify certain test scenarios where each different bandwidth method is particularly suitable (e.g., Wallin et al., 2018). Therefore, by considering that it is reasonable to claim that selection of bandwidth could have a noteworthy role in the performance of Kernel equating methods, the present study examines the role of bandwidth selection on Kernel test equating methods' SEEs on TIMSS data.

Third, as an alternative to NEAT test design, relevant literature shows that test equating can be conducted by using NEC design (e.g., Albano & Wiberg, 2019; Akın Arıkan, 2020; Branberg, 2010; Branberg & Wiberg, 2011; Gonzales et al., 2015; Wiberg & Branberg, 2015; Wiberg & von Davier, 2017). Indeed, a number of studies revealed that the covariate variables can reduce the amount of equating error (e.g., Branberg & Wiberg, 2011) and covariates can be used when equating different tests (e.g., Gonzales et al., 2015). Furthermore, in some studies, results showed that common variables can be used instead of common items. For example, in Yurtçu's (2018) study, scores were equated with gender and mathematics self-efficacy scores as covariates and common item scores and the study results showed that common variables can be used instead of common items. Akın Arıkan (2020) made a comparison with the NEAT design and NEC design using gender and socioeconomic status variables as covariance variables and concluded that in the absence of anchor tests, equating can be made by using common variables. Therefore, it may be argued that using covariates instead of common items in test equating may have a role in SEEs when equating scores are obtained from different TIMSS booklets. Hence, examining the role of test equating methods and bandwidths by taking into account test equating designs (i.e., NEAT and NEC) on SEEs when equating scores obtained from different TIMSS booklets is important to decide the eligible test equating approach.

2. METHOD

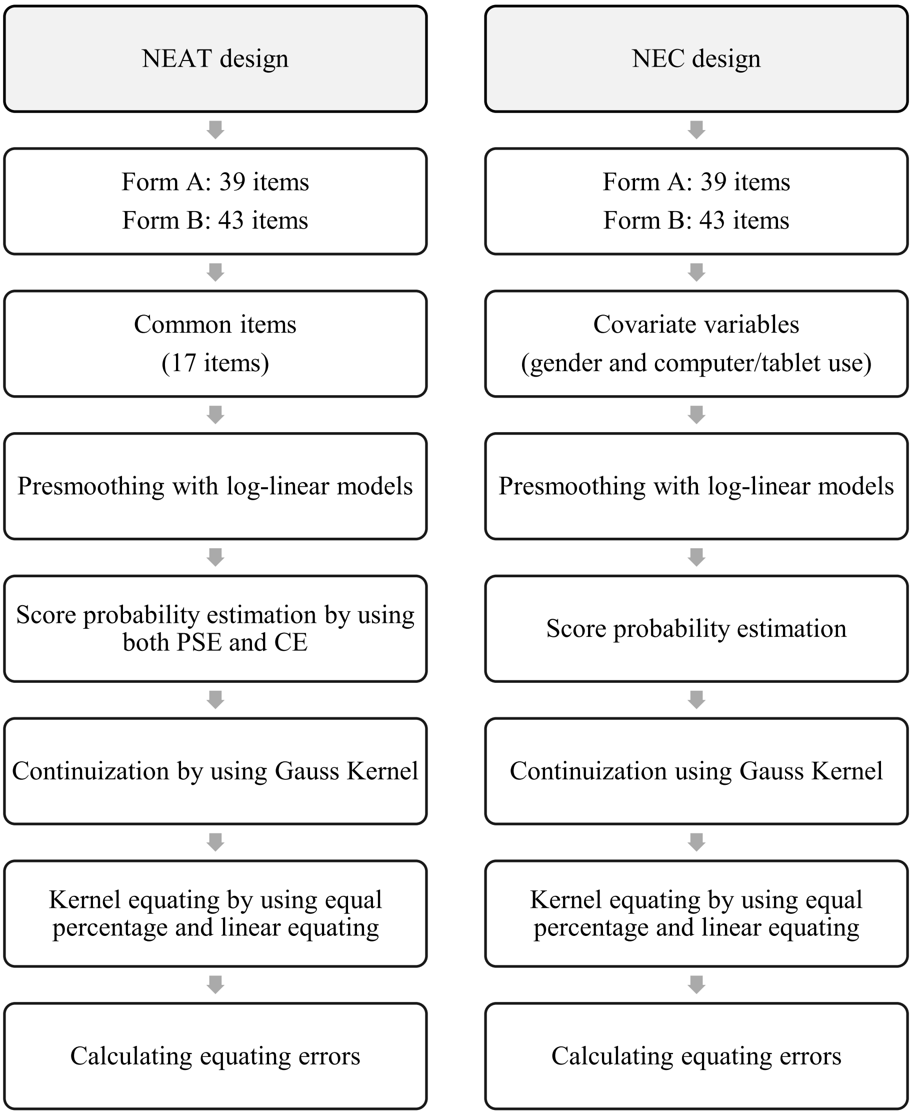
2.1. Study Group

In the study, two booklets numbered 1 and 14 out of 14 different booklets used in the Türkiye sample of the TIMSS 8th grade science test applied in 2019 were included in the analysis. 288 and 295 students took the specified tests, respectively. However, those students who did not answer the items in the student questionnaire were not included in the analysis. Therefore, the study group of the research consisted of 577 students, of whom 284 answered booklet number 1 and 293 answered booklet number 14.

2.2. Procedure

In the study, NEAT and NEC designs were used for Kernel test equating. The first booklet has 43 items, 17 of which are common, and the 14th booklet has 39 items, 17 of which are common. Science data belonging to booklets numbered 1 and 14 were converted into items with double scores as 1-0. For this, correct, partial credit, and full credit answers were coded as 1 point, while blank or wrong answers were coded as 0 point. While the gender variable was coded as 1=Girl and 2=Male, the computer/tablet variable was coded as 1=Yes and 2=No. In addition, gender and computer/tablet use variables were used as covariates in the NEC design in this study (see Figure 3).

Figure 3. *Research schema.*


NEAT design
Form A: 39 items
Form B: 43 items
Common items
(17 items)
Presmoothing with log-linear models
Score probability estimation by using both PSE and CE
Continuization by using Gauss Kernel
Kernel equating by using equal percentage and linear equating
Calculating equating errors
NEC design
Form A: 39 items
Form B: 43 items
Covariate variables
(gender and computer/tablet use)
Presmoothing with log-linear models
Score probability estimation
Continuization using Gauss Kernel
Kernel equating by using equal percentage and linear equating
Calculating equating errors

For test equating methods in both NEAT and NEC designs, in the first stage, the datasets of the two groups were smoothed with log-linear models. In the second stage, the score probability distributions were estimated using the smoothed score distributions obtained in the first stage. At this stage, score probability estimation was made by means of chain and post-stratification equating in the NEAT design. Chain equating starts by creating two separate single group patterns. Then, the first test form is linked to the common items, and from the common items to the other test form. In the post-stratification equating, the two groups are combined to form the target population. In the post-stratification equating, marginal distributions in the target population were obtained for the two test forms. In the third stage, the continuation stage, the Gaussian Kernel was used to make the discrete score distributions continuous in both NEAT and NEC designs. In the fourth stage, the tests were equalized by using the optimal and large bandwidths between the score distributions that became continuous in both NEAT and NEC designs. Finally, the SEE value was calculated.

In sum, in the current study, ten Kernel test equating methods were compared under NEAT and NEC designs. Kernel test equating methods used in the NEAT design are Kernel post-stratification equating with optimal bandwidth, Kernel post-stratification with large bandwidth, Kernel chain equating with optimal bandwidth, and Kernel chain equating with large bandwidth. Kernel equating methods used in the NEC design are Kernel equating with optimal bandwidth using gender as a covariate, Kernel equating with optimal bandwidth using computer/tablet use as a covariate, Kernel equating with large bandwidth using gender as a covariate, Kernel equating with large bandwidth using computer/tablet use as a covariate, Kernel equating with optimal bandwidth using both gender and computer/tablet use as covariates, and lastly Kernel equating with large bandwidth using both gender and computer/tablet use as covariates (see Table 1).

Table 1. *Ten different Kernel test equating methods compared in the present study.*

		NEAT design		Bandwidth		NEC design	
		Chain equating	Post-stratification equating	Optimal	Large	Gender	Computer/ tablet use
NEAT design	Chain equating with optimal bandwidth under NEAT design	x		x			
	Chain equating with large bandwidth under NEAT design	x			x		
	Post-stratification equating with optimal bandwidth under NEAT design		x	x			
	Post-stratification equating with large bandwidth under NEAT design		x		x		
NEC design	Equating with optimal bandwidth using gender as covariate under NEC design			x		x	
	Equating with large bandwidth using gender as covariate under NEC design				x	x	
	Equating with optimal bandwidth using computer/tablet use as covariate under NEC design			x			x
	Equating with large bandwidth using computer/tablet use as covariate under NEC design				x		x
	Equating with optimal bandwidth using gender and computer/tablet use as covariates under NEC design			x		x	x
	Equating with large bandwidth using gender and computer/tablet use as covariates under NEC design				x	x	x

2.3. Data Analysis

In this specific research, the performance of Kernel test equating and bandwidth selection was examined under two different test equating designs (i.e., NEAT and NEC designs). Reliability coefficients and descriptive statistics for test forms were calculated using SPSS software before the analysis for equating. The *kequate* package (Andersson et al., 2013) was used through the R program (R Core Team, 2013) to equate the two test forms using kernel equation methods. Equation methods were compared using standard equation errors (SEE).

3. RESULT

In this study, Kernel test equating methods were compared under NEAT and NEC designs. Kernel test equating methods used in the NEAT design are Kernel post-stratification equating with optimal bandwidth, Kernel post-stratification with large bandwidth, Kernel chain equating with optimal bandwidth, and Kernel chain equating with large bandwidth. Kernel equating methods used in the NEC design are Kernel equating with optimal bandwidth using gender as a covariate, Kernel equating with optimal bandwidth using computer/tablet use as a covariate, Kernel equating with large bandwidth using gender as a covariate, Kernel equating with large bandwidth using computer/tablet use as a covariate, Kernel equating with optimal bandwidth using both gender and computer/tablet use as covariates, and lastly Kernel equating with large bandwidth using both gender and computer/tablet use as covariates. In the present study, data obtained from two booklets (i.e., booklet 1 and booklet 14) of TIMSS 2019 8th grade science test were used. Under these conditions, which Kernel test equating method/s gave less incorrect results was examined by comparing the SEEs.

3.1. Preliminary Analysis Results

Table 2 shows the means, standard deviations, and reliability scores of two test forms. When these results are examined, it can be seen that these two groups have similar means and similar standard deviations. Furthermore, it is also seen that test forms used in this study have high-reliability coefficients.

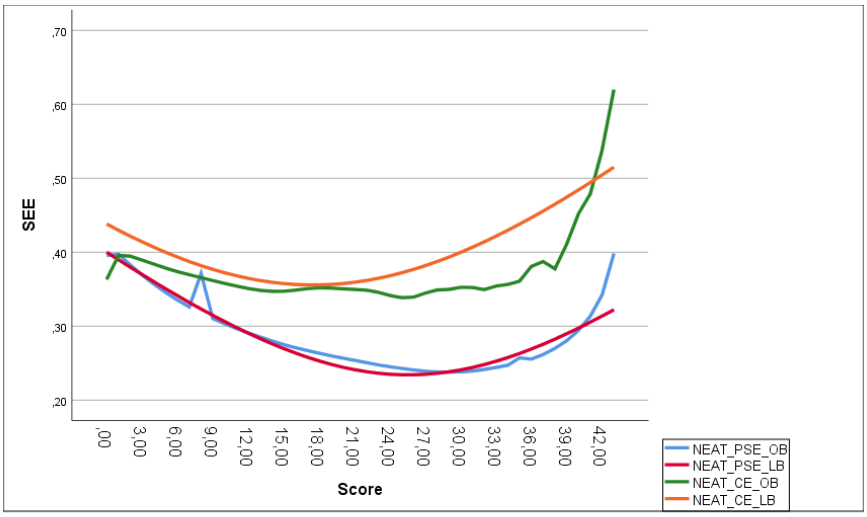
Table 2. *Descriptive statistics and reliability results of test forms.*

	Group 1 (n=284)		Group 2 (n=293)	
	Form A	Anchor items	Form B	Anchor items
Mean	24.25	10.04	22.06	9.48
St. Deviation	9.47	3.68	7.57	3.99
Skewness	-0.16	-0.28	-0.15	-0.23
Kurtosis	-1.00	-0.54	-0.84	-0.88
KR-20	.91	.76	.87	.84

3.2. Comparison of Kernel Test Equating Methods in the NEAT Design

When the standard error of the equating obtained as a result of the equating in the NEAT design is examined in Figure 4, it can be seen that the equating errors are similar for the low scores obtained from the tests. In general, regardless of bandwidth selection, Kernel post-stratification equation methods have been found to give lower error than that of chain equating methods. Kernel post-stratification equating methods show a similar distribution in terms of bandwidth. Although Kernel chain equating methods initially show a similar distribution, they differ in high scores. Specifically, Kernel chain equating with large bandwidth gives the highest SEEs for high scores. On the other hand, equalized scores obtained as a result of equating with the NEAT design are given in Table 5 in the appendices.

Figure 4. *Comparison of Kernel post-stratification and chain equating methods under NEAT design.*



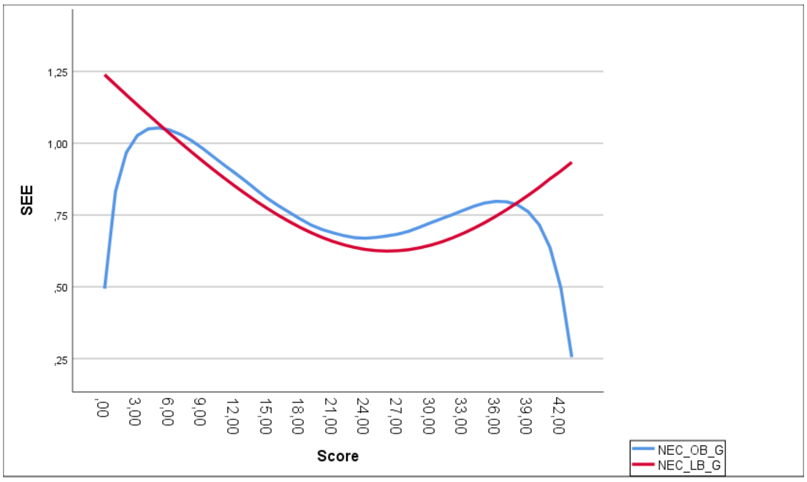
Note. PSE_OB = Post-stratification method using optimal bandwidth, PSE_LB = Post-stratification method using large bandwidth, CE_OB = Chain equating method using optimal bandwidth, CE_LB = Chain equating method using large bandwidth.

3.3. Comparison of Kernel Test Equating Methods in The NEC Design

3.3.1. *Gender as a covariate*

Similar results were observed in the Kernel equating methods using optimal and large bandwidths and gender as a covariate under the NEC design (see Figure 5). As can be seen in Figure 5, Kernel equating method using large bandwidth gives higher SEEs at the scores at the bottom and top of the test. On the other hand, Kernel equating method using optimal bandwidth gives the lowest SEEs in the scores at the upper and lower parts of the scale. Both Kernel test equating methods have similar error values in the middle parts of the scale. Additionally, equalized scores obtained as a result of equating with gender variable as a covariate under the NEC design are given in Table 6 in the appendices.

Figure 5. *Comparison of Kernel using gender as a covariate under NEC design.*

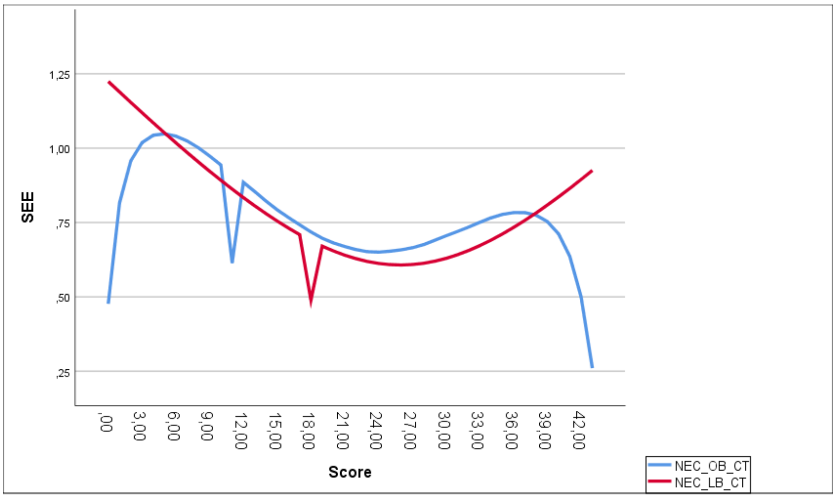


Note. OB_G= Equating method using optimal bandwidth by gender covariate, LB_G= Equating method using large bandwidth by gender covariate.

3.3.2. *Computer/tablet use as a covariate*

Figure 6 shows that Kernel equating method using large bandwidth and computer/tablet use as a covariate gives higher SEEs in the top and bottom of the scale, while Kernel equating method using optimal bandwidth and computer/tablet use as a covariate gives lower SEEs in the top and bottom of the scale. Both Kernel test equating methods have similar SEEs in the middle part of the scale. On the other hand, equalized scores obtained as a result of equating with computer/tablet use variable as a covariate under the NEC design are given in Table 7 in the appendices.

Figure 6. *Comparison of Kernel using computer/tablet use as a covariate under NEC design.*

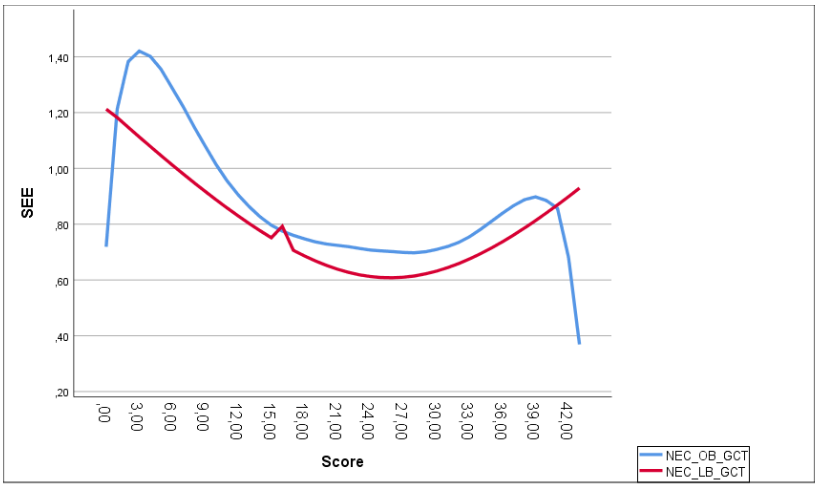


Note. OB_G= Equating method using optimal bandwidth by the use of computer/tablet covariate, LB_G= Equating method using large bandwidth by the use of computer/tablet covariate.

3.3.3. *Gender and computer/tablet use as covariates together*

Similar to the results related to previous variables, in the condition in which gender and computer/tablet are used as covariates together, Kernel equating method using large bandwidth gives higher SEEs in the scores at the bottom and top of the scale. On the other hand, Kernel equating using optimal bandwidth gives the lower error in the scores in the upper and lower parts of the scale. Both Kernel test equating methods have similar SEEs in the middle parts of the scale (see Figure 7). Besides these similar results, it can be seen that Kernel equating method under NEC design in which gender and computer/tablet variables are included together have higher SEE values than the SEEs obtained from Kernel test equating methods in which these variables were considered separately. Additionally, equalized scores obtained as a result of equating with gender and computer/tablet use variables as covariates under the NEC design are given in Table 8 in the appendices.

Figure 7. *Comparison of Kernel using gender and computer/tablet use as covariates together.*

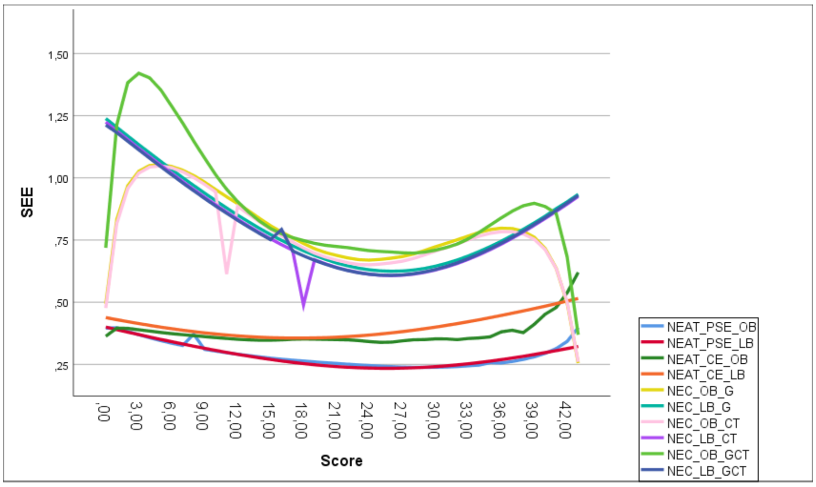


Note. OB_G= Equating method using optimal bandwidth by gender and the use of computer/tablet covariates, LB_G= Equating method using large bandwidth by gender and the use of computer/tablet covariates.

3.4. Comparison of the Role of Bandwidth Selection in SEEs in Both NEAT and NEC Design

In the current study, optimal and large bandwidths were determined by kequate R package. The bandwidths for the Kernel post-stratification equating method are h(X) = 0.49 and h(Y) = 0.65. For the Kernel post-stratification equating method, the large bandwidths are h(X) = 12694.88 and h(Y) = 4190.42 (see Table 3). The results show that Kernel post-stratification methods using optimal and large bandwidth under the NEAT design demonstrate similar results. Furthermore, these methods demonstrate the lowest SEEs in the NEAT design (see Figure 8).

Figure 8. *Comparison of ten Kernel test equating methods in terms of bandwidth selection.*



Note. PSE_OB = Post-stratification method using optimal bandwidth, PSE_LB = Post-stratification method using large bandwidth, CE_OB = Chain equating method using optimal bandwidth, CE_LB = Chain equating method using large bandwidth.

In the Kernel chained equating, two linking functions, from X to A on P (group answering form X) and from A to Y on Q (group answering form Y) were used. Therefore, four distributions were to be continuized (von Davier et al., 2006). The optimal bandwidths are h(X) = 0.49, h(AP)= 0.43 and h(Y) = 0.68, h(AQ) = 0.43. The large bandwidths of the same equating method are h(X) = 12662.04, h(AP)= 3524.42 and h(Y) = 4183.10, h(AQ) = 5481.53 (see Table 3). Although Kernel chain equating methods with both optimal and large bandwidths in NEAT design gave higher error than that of Kernel post-stratification equating methods under NEAT design, they resulted in fewer errors than all Kernel equating methods under NEC design. Although Kernel chain equating methods using optimal and large bandwidths initially showed a similar distribution, they differed in equating high scores. Specifically, Kernel chain equating using large bandwidth gave the highest SEEs for high scores. Although Kernel equating methods using optimal bandwidth gave fewer SEEs in scores at the top and bottom of the scale, in general Kernel equating methods using large bandwidth demonstrated fewer SEEs (see Figure 8).

Table 3. *Bandwidths (h parameters) for NEAT design.*

	PSE-OB	PSE-LB	CE-OB	CE-LB
h_x	0.49	12694.88	0.49	12662.04
h_y	0.65	4190.42	0.68	4183.10
h_{aP}			0.43	3524.42
h_{aQ}			0.45	5481.53

Note. PSE-OB = Post-stratification method using optimal bandwidth, PSE-LB = Post-stratification method using large bandwidth, CE-OB = Chain equating method using optimal bandwidth, CE-LB = Chain equating method using large bandwidth.

In NEC design, the optimal bandwidths for the Kernel equating method in which gender was used as covariate are h(X) = 0.56 and h(Y) = 0.60. For the same equating method, the large bandwidths are h(X) = 9773.86 and h(Y) = 7563.13. The optimal bandwidths for the Kernel equating method in which computer tablet was used as covariate are h(X) = 0.56 and h(Y) = 0.60. For the same equating method, the large bandwidths are h(X) = 9474.77 and h(Y) = 7559.21. The optimal bandwidths for the Kernel equating method in which gender and computer/tablet use were applied as covariate are h(X) = 0.56 and h(Y) = 0.60. For the same equating method, the large bandwidths are h(X) = 9522.01 and h(Y) = 7633.00 (see Table 4). In the NEC design for all covariate options, although they gave higher SEEs for the scores at the top and bottom of the scale, the methods using large bandwidth gave less error regardless of covariate selection.

Table 4. *Bandwidth (h parameters) for NEC design.*

	G-OB	G-LB	CTU-OB	CTU-LB	G&CTU-OB	G&CTU-OB
h_x	0.56	9473.86	0.56	9474.77	0.56	9522.01
h_y	0.60	7563.13	0.60	7559.21	0.60	7633.00

Note. G-OB = Gender-optimal bandwidth, G-LB = Gender-large bandwidth, CTU-OB = Computer/tablet use-optimal bandwidth, CTU-LB = Computer/tablet use-large bandwidth, G&CTU-OB = Gender and computer/tablet use-optimal bandwidth, G&CTU-LB = Gender and computer/tablet use-large bandwidth.

In sum, NEAT design demonstrated lowest SEE values compared to those of NEC design, regardless of the covariate variable/s and bandwidth used. Kernel post-stratification equation methods showed a similar distribution in terms of bandwidth. Although Kernel chain equating methods initially showed a similar distribution, they differed in high scores. Specifically, Kernel chain equating with large bandwidth gave the highest SEEs for high scores. Kernel post-stratification equating methods resulted in less SEE than that of Kernel chain equating methods. NEC design was the design with the highest SEE values overall, regardless of the covariate variable/s and bandwidth used. When the methods based on NEC design are evaluated based on bandwidth in themselves, for all covariate options, the methods using large bandwidth gave less error. When the methods based on the NEC design were evaluated in terms of covariate selection, it was seen that the test equating methods in which gender and computer/tablet variables were handled separately resulted in less SEE than those in which these variables were considered together.

4. DISCUSSION and CONCLUSION

In this study, Kernel test equating methods were compared under NEAT and NEC designs. In NEAT design, taking into account optimal and large bandwidths, Kernel post-stratification and chain equating methods were compared. In the NEC design, gender and/or computer/tablet use was considered as a covariate, and by using these covariates Kernel test equating methods were performed. In these comparisons, bandwidths were considered as well.

In research that compares performance or errors of different methods, the main question asked is which methods should be preferred. In line with previous research (e.g., Akın Arıkan, 2019), the current study has shown that the Kernel post-stratification equating method provides fewer SEEs than those of the Kernel chain equating method in NEAT design. However, some studies show that the post-stratification method was more biased than the chain equating method (e.g., Livingston et al., 1990). It should be noted that these studies emphasize that the chain equating method may be preferable to post-stratification equating methods when the groups differ widely on the anchor test. In the current study, the reason why post-stratification methods show less error compared to chain equating may be that the two groups in this study have similar achievements.

The general finding of the current study is that Kernel equating methods in NEAT design resulted in fewer errors than those of Kernel equating methods in NEC design. The current study is consistent with previous studies (e.g., Akın Arıkan 2020; Wiberg & Branberg, 2015) which showed that NEAT design provides more accurate results in comparison to NEC design. Given that the performance of NEC design in equating depends on how well the covariates predict test scores and how well background variables explain differences in test scores (Wiberg & Branberg, 2015), one of the possible explanations for these results can be the selection of covariates. Gender and computer tablet use may not be eligible covariates for this group or this discipline (i.e., science). On the other hand, the second possible explanation may be sample size. In this study, Kernel equating was performed under the NEC design using a small sample; however, use of a small sample size can have caused the risk of having sparse data in some cells to be increased. Indeed, related studies that take into account the sample size in the NEC design (e.g., Branberg & Wiberg, 2011; Gonzales et al., 2015) show that the equating errors of the equating using the covariate under the NEC design are higher in the small sample. These explanations are also valid for results that show that conditions in which gender and computer/tablet use variables were handled together resulted in more SEEs in comparison to conditions in which these variables were handled separately in NEC design. As one adds more covariates, one obtains a rapid increase in the number of categories. Given that adding more covariates increases the risk of having sparse data in some cells (Wiberg & Branberg, 2015), it can be understood why conditions in which gender and computer/tablet use variables were handled together resulted in more SEEs in comparison to conditions in which these variables were handled separately in NEC design. Although the results of this study show that Kernel test equating methods in NEAT design give fewer SEEs, compared to SEEs of NEC design in which covariates (i.e., gender and computer/tablet use) were used, they still provide useful information to the literature on test equating. If even more suitable covariates for the distribution of the groups could be found and research could be replicated in large samples, equating performance could be closer to the results obtained with the NEAT design. At this point, more research is needed on NEC design.

Based on the results of the current study, five main conclusions can be drawn as follows:

1) In the NEAT design, Kernel chain equating methods (for both optimal and large bandwidths) exhibited higher error than the post-stratification equating methods did (for both optimal and large bandwidths).

2) The lowest error in the NEC design was obtained from the Kernel equating method with large bandwidth through the computer/tablet variable.

3) Kernel test equating results based on the NEC design, which considers gender and computer tablet use variables as a covariate separately, showed lower SEE than that of the NEC pattern, which takes these variables together as covariates.

4) In terms of bandwidth, when all methods are compared within the pattern used (i.e., NEAT and NEC), it has been seen that generally Kernel test equating results with large bandwidth result in fewer errors than the Kernel test equating results with optimal bandwidth.

5) When the NEAT and NEC designs are compared generally, the NEAT design has a lower SEE than that of the NEC design.

The results of the current study showed that in the NEAT design, Kernel post-stratification equating methods give fewer SEEs compared to those of Kernel chain equating methods. Given that relatively small samples were used in the present study, it can be recommended that in studies to be conducted on small samples, post-stratification equating methods be preferred to Kernel chain equating methods. The present study results also showed that the Kernel test equating results with large bandwidth result in fewer errors than the Kernel test equating results with optimal bandwidth. In test equating studies on small samples, large bandwidth can be preferred.

In the present study, Kernel test equating results based on the NEC design, which considers gender and computer tablet use variables as a covariate separately, showed lower SEE than that of the NEC design, which takes these variables together as covariates. In practice, using more than one covariate could be a reason for the inflation of SEEs because of increase in category number. Therefore, if covariates are going to be used in test equating studies that do not use anchor tests for equating, it should be noted that covariate number can be a reason for the increase in SEEs.

Lastly and most importantly, the results of the study showed that Kernel equating methods using anchor tests give fewer SEEs compared to those using covariate/s. In practice, it could be recommended that researchers or educators prefer to apply anchor tests instead of covariates. However, the present study did not address using anchor tests or students' demographic variables together as covariates. In further studies, those effects of usage of anchor tests as well as demographic variables as a covariate in NEC design can be examined.

Acknowledgments

The preliminary results of this study have been presented in 7th International Measurement and Evaluation in Education and Psychology Congress.

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

Seyma Nur Ozsoy: Conception, Design, Fundings, Materials, Data Collection and/or Processing, Analysis and/or Interpretation, Literature Review, Writing, Critical Review. **Sevilay Kilmen**: Supervision, Critical Review.

Orcid

Seyma Nur Ozsoy https://orcid.org/0000-0002-8306-6114
Sevilay Kilmen https://orcid.org/0000-0002-5432-7338

REFERENCES

Akın Arıkan, Ç. (2020). The impact of covariate variables on kernel equating under the non-equivalent groups. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme, 11*(4), 362-373.

Albano, A.D., & Wiberg, M. (2019). Linking with external covariates: examining accuracy by anchor type, test length, ability difference, and sample size. *Applied Psychological Measurement*, *43*(8), 597-610.

Andersson, B., Bränberg, K., & Wiberg, M. (2013). Performing the kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1-25.

Bränberg, K. (2010). *Observed score equating with covariates* [Unpublished Doctoral dissertation]. Department of Statistics, Umeå University.

Branberg, K., & Wiberg, M. (2011). Observed score linear equating with covariates. *Journal of Educational Measurement, 48*(4), 419-440.

Braun, H.I., & Holland, P.W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P.W. Holland & D.B. Rubin (Eds.) *Test equating* (9-49). Academic Press.

Choi, S.I. (2009). *A comparison of kernel equating and traditional equipercentile equating methods and the parametric bootstrap methods for estimating standard errors in* *equipercentile equating* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Godfrey, K.E. (2007). *A comparison of kernel equating and IRT true score equating methods* [Unpublished doctoral dissertation]. The Faculty of the Graduate School at the University of North Carolina at Greensboro.

Gonzales, J., Barrientos, A.F., & Quintana, F.A. (2015). Bayesian nonparametric estimation of test equating functions with covariates. *Computational Statistics and Data Analysis, 89*, 222-244.

Gonzales, J., & Wiberg, M. (2017). *Applying test equating methods*. Springer.

Kolen, M.J. (1988). Traditional equating methodology. *Educational measurement: Issues and Practice*, *7*(4), 29-37.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practises*. Springer

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling and linking: Methods and practises.* Springer

Liang, T., & von Davier, A.A. (2014). Cross-validation: An alternative bandwidth-selection method in Kernel equating. *Applied Psychological Measurement, 38*(4), 281-295.

Livingston, S.A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

Livingston, S.A. (2014). Equating test scores (without IRT). *Educational testing service.*

Livingston, S.A., Dorans, N.J., & Wright, N.K. (1990). What combination of sampling equating methods works best?. *Applied Measurement in Education, 3*, 73–95.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

R Core Team. (2013). *R: A language and environment for statistical computing.* (Versiyon 4.0.3) [Computer software]. R Foundation for Statistical Computing.

von Davier, A.A. (2013). Observed-score equatıng: an overvıew. *Psychometrika, 78(*4), 605-623.

von Davier, A.A., Holland, P.W., Livingston, S.A., Casabianca, J., Grant, M.C., & Martin, K. (2006). An evaluation of the kernel equating method: A special study with pseudo tests constructed from real test data. *ETS Research Report Series*, *2006*(1), i-31.

von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004a). The chain and post-stratification methods for observed-score equating: their relationship to population invariance. *Journal of Educational Measurement, 41*(1), 15-32.

von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004b). *The kernel method of test equating*. Springer.

von Davier, A.A., & Chen, H. (2013). The Kernel levine equipercentile observed-score equating function. *ETS Research Report Series*, *2013*(2), i-27.

Wallin G., Häggström J., & Wiberg M. (2018) *How to select the bandwidth in kernel equating-An evaluation of five different methods*. In Wiberg M., Culpepper S.,Janssen R., González J., & Molenaar D. (Ed.), *Quantitative Psychology*. IMPS 2017. Springer Proceedings in Mathematics & Statistics, vol 233. Springer. https://doi.org/10.1007/978-3-319-77249-3_8

Wiberg, M. (2015). A note on equating test scores with covariates. In Ellinor Fackle- Fornius (Ed), *Festschrift in Honor of Hans Nyquist on The Occasion of His 65th Birthday.* Stockholm University.

Wiberg, M., & Branberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39*(5), 349-361.

Wiberg, M., & von Davier, A.A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing, 17*(2), 105-126.

Yurtçu, M. (2018). *The comparison of test equating with covariates using Bayesian nonparametric method* [Unpublished master thesis]. Hacettepe University.

APPENDICES

Table 5. *Equalized scores obtained in the NEAT design.*

	NEAT_PSE_EQ	NEAT_PSE_L	NEAT_CE_EQ	NEAT_CE_L
1	2.26	4.55	1.45	3.44
2	3.92	4.88	2.71	3.65
3	4.89	5.21	3.43	3.86
4	5.60	5.54	3.95	4.08
5	6.19	5.87	4.38	4.29
6	6.70	6.20	4.74	4.50
7	7.15	6.53	5.07	4.71
8	7.57	6.86	5.37	4.93
9	7.94	7.19	5.64	5.13
10	8.28	7.52	5.89	5.35
11	8.60	7.85	6.12	5.56
12	8.91	8.18	6.34	5.78
13	9.21	8.51	6.53	5.99
14	9.50	8.84	6.71	6.20
15	9.78	9.17	6.88	6.41
16	10.04	9.50	7.05	6.63
17	10.29	9.83	7.21	6.84
18	10.51	10.16	7.38	7.05
19	10.74	10.49	7.54	7.26
20	10.97	10.82	7.69	7.48
21	11.20	11.15	7.85	7.69
22	11.43	11.48	8.01	7.90
23	11.68	11.81	8.17	8.11
24	11.93	12.14	8.34	8.32
25	12.19	12.47	8.50	8.54
26	12.45	12.80	8.66	8.75
27	12.71	13.13	8.83	8.96
28	12.97	13.46	9.01	9.17
29	13.23	13.79	9.19	9.38
30	13.51	14.12	9.38	9.60
31	13.80	14.45	9.57	9.81
32	14.10	14.78	9.77	10.02
33	14.41	15.11	9.97	10.24
34	14.72	15.44	10.17	10.45
35	15.04	15.77	10.38	10.66
36	15.35	16.10	10.61	10.87
37	15.69	16.43	10.85	11.09
38	16.06	16.76	11.13	11.30
39	16.47	17.09	11.42	11.51
40	16.95	17.42	11.75	11.72
41	17.53	17.75	12.24	11.94
42	18.26	18.08	12.83	12.15
43	19.27	18.41	13.81	12.36
44	21.09	18.74	15.93	12.57

Table 6. *Equalized scores obtained in the NEC design by the use computer/tablet covariate.*

	NEC_EQ_GK	NEC_L_GK	NEC_EQ_LK	NEC_EQ_UK
1	0.73	2.73	0.83	0.68
2	2.35	3.53	2.36	2.38
3	3.66	4.33	3.66	3.68
4	4.79	5.13	4.79	4.79
5	5.81	5.93	5.81	5.81
6	6.75	6.73	6.75	6.76
7	7.65	7.53	7.65	7.66
8	8.52	8.32	8.52	8.54
9	9.36	9.12	9.36	9.37
10	10.18	9.92	10.18	10.19
11	10.99	10.72	10.99	10.99
12	11.79	11.52	11.79	11.79
13	12.58	12.32	12.58	12.59
14	13.37	13.11	13.37	13.37
15	14.15	13.91	14.14	14.14
16	14.92	14.71	14.92	14.92
17	15.69	15.51	15.69	15.69
18	16.46	16.31	16.46	16.47
19	17.23	17.11	17.23	17.23
20	17.99	17.90	17.99	17.99
21	18.76	18.70	18.76	18.76
22	19.52	19.50	19.52	19.53
23	20.29	20.30	20.29	20.29
24	21.05	21.10	21.05	21.05
25	21.82	21.90	21.82	21.82
26	22.58	22.69	22.58	22.58
27	23.35	23.49	23.35	23.35
28	24.12	24.29	24.12	24.12
29	24.90	25.09	24.90	24.90
30	25.67	25.89	25.67	25.67
31	26.45	26.69	26.45	26.45
32	27.24	27.48	27.24	27.24
33	28.03	28.28	28.03	28.04
34	28.84	29.08	28.84	28.84
35	29.65	29.88	29.65	29.64
36	30.47	30.68	30.48	30.46
37	31.32	31.48	31.32	31.31
38	32.18	32.28	32.18	32.18
39	33.08	33.07	33.08	33.08
40	34.01	33.87	34.02	34.01
41	35.00	34.67	35.01	35.00
42	36.07	35.47	36.07	36.07
43	37.27	36.27	37.26	37.26
44	38.62	37.07	38.57	38.65

Table 7. *Equalized scores obtained in the NEC design by gender covariate.*

	NEC_EQ_GK	NEC_L_GK	NEC_EQ_LK	NEC_EQ_UK
1	0.78	2.79	0.88	0.73
2	2.43	3.59	2.43	2.46
3	3.75	4.39	3.74	3.76
4	4.87	5.19	4.87	4.88
5	5.89	5.99	5.89	5.89
6	6.84	6.78	6.83	6.84
7	7.73	7.58	7.73	7.74
8	8.60	8.38	8.59	8.61
9	9.43	9.18	9.43	9.45
10	10.25	9.98	10.25	10.26
11	11.06	10.78	11.06	11.06
12	11.86	11.57	11.86	11.86
13	12.64	12.37	12.64	12.65
14	13.42	13.17	13.42	13.44
15	14.20	13.97	14.20	14.20
16	14.97	14.78	14.97	14.97
17	15.74	15.57	15.74	15.74
18	16.51	16.36	16.51	16.52
19	17.28	17.16	17.28	17.28
20	18.04	17.96	18.04	18.04
21	18.80	18.76	18.80	18.80
22	19.57	19.56	19.57	19.57
23	20.33	20.36	20.33	20.33
24	21.10	21.15	21.10	21.10
25	21.86	21.95	21.86	21.86
26	22.63	22.75	22.63	22.63
27	23.40	23.55	23.40	23.40
28	24.17	24.35	24.17	24.17
29	24.95	25.15	24.95	24.95
30	25.73	25.94	25.73	25.72
31	26.51	26.74	26.51	26.50
32	27.30	27.54	27.30	27.29
33	28.09	28.34	28.09	28.09
34	28.90	29.14	28.90	28.90
35	29.71	29.94	29.71	29.71
36	30.54	30.73	30.54	30.53
37	31.39	31.53	31.39	31.38
38	32.26	32.33	32.26	32.25
39	33.15	33.13	33.15	33.15
40	34.09	33.93	34.09	34.09
41	35.08	34.73	35.08	35.08
42	36.14	35.52	36.14	36.14
43	37.32	36.32	37.32	37.31
44	38.65	37.12	38.60	38.68

Table 8. *Equalized scores obtained in the NEC design by gender and the use computer/tablet covariate.*

	NEC_EQ_GK	NEC_L_GK	NEC_EQ_LK	NEC_EQ_UK
1	0.66	2.85	0.75	0.62
2	2.23	3.65	2.24	2.24
3	3.54	4.44	3.54	3.56
4	4.67	5.24	4.67	4.68
5	5.70	5.79	5.70	5.71
6	6.67	6.84	6.67	6.68
7	7.59	7.63	7.59	7.61
8	8.48	8.43	8.48	8.50
9	9.34	9.23	9.34	9.35
10	10.19	10.03	10.19	10.19
11	11.02	10.83	11.02	11.02
12	11.84	11.62	11.84	11.84
13	12.65	12.42	12.65	12.66
14	13.45	13.22	13.45	13.46
15	14.24	14.02	14.24	14.25
16	15.03	14.82	15.03	15.03
17	15.82	15.61	15.82	15.82
18	16.60	16.41	16.60	16.60
19	17.38	17.21	17.37	17.38
20	18.15	18.01	18.15	18.15
21	18.92	18.80	18.92	18.92
22	19.69	19.60	19.69	19.69
23	20.45	20.40	20.45	20.46
24	21.22	21.20	21.22	21.22
25	21.98	22.00	21.98	21.98
26	22.75	22.79	22.75	22.75
27	23.51	23.59	23.51	23.51
28	24.27	24.39	24.27	24.27
29	25.04	25.19	25.04	25.04
30	25.80	25.98	25.80	25.80
31	26.57	26.78	26.57	26.57
32	27.35	27.58	27.35	27.34
33	28.13	28.38	28.13	28.13
34	28.91	29.18	28.91	28.91
35	29.71	29.97	29.71	29.70
36	30.52	30.77	30.52	30.50
37	31.34	31.57	31.34	31.34
38	32.19	32.37	32.19	32.19
39	33.07	33.17	33.07	33.07
40	33.99	33.96	33.99	33.99
41	34.97	34.76	34.97	34.97
42	36.03	35.56	36.03	36.03
43	37.22	36.36	37.22	37.21
44	38.60	37.15	38.54	38.62

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 76–97

https://doi.org/10.21449/ijate.1102435

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

Spatial ability test for university students: Development, validity and reliability studies

Kubra Acikgul[1,*], **Suleyman Nihat Sad**[2], **Bilal Altay**[2]

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

ARTICLE HISTORY

Received: Apr. 12, 2022

Accepted: Feb. 26, 2023

Keywords:

Student evaluation,
Test development,
Spatial ability.

Abstract: This study aimed to develop a useful test to measure university students' spatial abilities validly and reliably. Following a sequential explanatory mixed methods research design, first, qualitative methods were used to develop the trial items for the test; next, the psychometric properties of the test were analyzed through quantitative methods using data obtained from 456 university students. As a result, a multiple-choice spatial ability test with 27 items and five options was created, divided into three subtests: spatial relations, spatial visualization, and spatial orientation. The results suggested that scores obtained from the spatial ability test and its subtests are valid and reliable.

1. INTRODUCTION

Spatial ability is regarded as a critical component of human abilities (Lohman, 1993) and a prerequisite for scientific thinking (Clements & Battista, 1992). Spatial ability has an important role in the assimilation and use of preexisting knowledge as well as in the development of new knowledge and creativity (Kell et al., 2013). For example, the mental rotation skill is apparently an inevitable spatial ability for some popular professions, including dentistry, medicine, architecture, interior design, engineering, navigation, etc. (Kerkman et al., 2000). The decisive role of spatial abilities in the development of knowledge and skills in the fields of science, technology, engineering, and mathematics (STEM) is emphasized in many studies (e.g., Contreras et al., 2018; Gilligan et al., 2017). More specifically, spatial abilities are reported to have a critical role in enhancing the performance of learning mathematics and geometry (Battista et al., 1982; Gilligan et al., 2017; Sarama & Clements, 2009) and in developing mathematical thinking skills (Young et al., 2018).

Despite the importance and key role attributed to spatial abilities in many fields, the lack of a clear consensus on the definition and components of spatial ability confuses measuring spatial ability (D'Oliveira, 2004; Eliot & Hauptman, 1981; National Research Council (NRC), 2006). D'Oliveira (2004) reviewed the four main reasons for this confusion as follows: 1) Different definitions ascribed to spatial ability, 2) Different numbers of components of spatial ability, 3)

*CONTACT: Kübra AÇIKGÜL ✉ kubra.acikgul@inonu.edu.tr İnönü University, Faculty of Education, Department of Mathematics and Science Education, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

Different names given to the components of spatial ability, and 4) Quite a variety of spatial ability tests. In a similar vein, Eliot and Hauptman (1981) asserted that inconsistency among the methods and tools used to measure spatial ability further complicated the problem of a lack of consensus in the spatial ability literature. Thus, this study aimed to present a detailed review of the literature on measuring spatial ability first and then to develop a spatial ability test to include spatial relations, spatial visualization, and spatial orientation factors, which can measure the spatial abilities of university students in a valid, reliable, and useful way.

1.1. Literature Review

1.1.1. *Spatial ability and its components*

Spatial ability research started in the late 1800s with studies aimed at demonstrating that spatial ability is a separate factor from general intelligence and continued with studies aimed at identifying and defining the composition of spatial ability (Mohler, 2008). Since the concept of spatial ability was first introduced, many terms, including spatial ability, spatial reasoning, spatial concepts, spatial intelligence, spatial cognition, mental maps, environmental cognition, and cognitive mapping, have been used interchangeably in the spatial ability literature (NRC, 2006), and the term "spatial ability" has been defined in different ways (D'Oliveira, 2004; Martín-Dorta et al., 2008). In the present study, the concept of spatial ability was preferred since it is used more frequently in the field.

Gardner (1983), one of the leading theorists who popularized spatial ability with a different name, namely spatial intelligence, defined it as "the capacities to perceive the visual world accurately, to perform transformations and modifications upon one's initial perceptions, and to be able to re-create aspects of one's visual experience, even in the absence of relevant physical stimuli." (p. 173). Linn and Petersen (1985) defined spatial ability as a “skill in representing, transforming, generating, and recalling symbolic, nonlinguistic information.” (p.1482). Lohman (1993) defined it as “the ability to generate, retain, retrieve, and transform well-structured visual images.” (p.3). According to the National Research Council (2006), "spatial ability" is " a trait that a person has and as a way of characterizing a person’s ability to perform mentally such operations as rotation, perspective change, and so forth." (p.26). Tartre (1990) defined spatial skills as “mental skills concerned with understanding, manipulating, reorganizing, or interpreting relationships visually.” (p.216). According to Carroll (1993), individuals' spatial and other visual abilities refer to “searching the visual field, apprehending the forms, shapes, and positions of objects as visually perceived, forming mental representations of those forms, shapes, and positions, and manipulating such representations mentally." (p.304). Based on these definitions, the present study defines spatial ability as the ability to generate, retain, retrieve, manipulate, interpret, and reorganize the mental representations of visual objects by perceiving their forms and positions.

Psychometric studies on spatial ability have indicated that spatial ability does not have a monolithic structure, but is made up of a composition of factors consisting of sub-skills (Lohman, 1979, 1993; Guilford et al., 1952; Mohler, 2008). D'Oliveira (2004) also reported that one should refer to a domain of spatial abilities instead of a single spatial ability. Lohman (1993) noted that there are several spatial abilities, each focusing on different processes such as generating, retaining, retrieving, and transforming images. On the other hand, there are remarkable discrepancies and confusion in terms of the number and naming of factors in the literature and in terms of the tests used to measure each factor (D'Oliveira, 2004; Martín-Dorta et al., 2008). Relevant literature typically classifies spatial ability under two (Clements, 1998; Guilford et al., 1952; McGee, 1979; Pellegrino et al., 1984), three (Barnea, 2000; Contero et al., 2005; D'Oliveira, 2004; Linn & Petersen, 1985; Lohman, 1979); or five (Carroll, 1993; Maier, 1996) factors. These factors are summarized in Table 1.

Table 1. *The names of the factors classified by different researchers.*

Factors Author(s)	(Spatial) Visualization	Spatial Relations	(Spatial) Orientation	Spatial Perception	Mental Rotation	Perceptual Speed	Closure Speed	Flexibility of Closure
Clements (1998)	✓		✓					
McGee (1979)	✓		✓					
Pellegrino et al. (1984)	✓	✓						
Guilford et al. (1952)	✓	✓						
Linn and Petersen (1985)	✓			✓	✓			
Lohman (1979)	✓	✓	✓					
Barnea (2000)	✓	✓	✓					
Contero et al. (2005)	✓	✓	✓					
D'Oliveira (2004)	✓	✓	✓					
Maier (1996)	✓	✓	✓	✓	✓			
Carroll (1993)	✓	✓				✓	✓	✓

When these classifications are examined, it can be said that the most commonly-mentioned factors of spatial ability in the literature are Spatial Relations (Barnea, 2000; Carroll, 1993; Contero et al., 2005; D'Oliveira, 2004; Guilford et al., 1952; Lohman, 1979; Maier, 1996; Pellegrino et al., 1984), Spatial Visualization (Barnea, 2000; Carroll, 1993; Clements, 1998; Contero et al., 2005; D'Oliveira, 2004; Guilford et al., 1952; Linn & Petersen, 1985; Lohman, 1979; Maier, 1996; McGee, 1979; Pellegrino et al., 1984) and Spatial Orientation (Barnea, 2000; Clements, 1998; Contero et al., 2005; D'Oliveira, 2004; Lohman, 1979; Maier, 1996; McGee, 1979). Therefore, it was decided to include spatial visualization, spatial relations, and spatial orientation factors as components of the spatial ability test developed in this study.

1.1.2. *Measuring spatial ability*

Different definitions of spatial ability and its components have led to the use of many different tests for measuring these abilities (Martín-Dorta et al., 2008). There is a large variety of spatial ability tests, which confuses their names and content (D'Oliveira, 2004).

Table 2 below provides various definitions of spatial visualization, spatial relations, and spatial orientation factors, which allow the readers to examine the definitions more clearly in a comparative manner and the most popular tests measuring the different components of spatial ability.

As can be seen from the definitions above, the confusion in defining spatial ability is also true for its components. To illustrate, both spatial visualization and spatial relations abilities are defined in terms of mental rotation ability. This situation also confuses the measurement of these abilities, as many researchers have stated (e.g., Carroll, 1993; D'Oliveira, 2004; Eliot & Hauptman, 1981). As a result, distinguishing the differences between spatial visualization and spatial relations abilities, as well as the types of items to be used for their measurement, is critical.

Table 2. *Some tests measuring the different components of spatial ability.*

	Components of spatial ability		
	Spatial visualization	Spatial relations	Spatial orientation
Definitions	"the ability to manipulate visual objects mentally.", (Guilford et al., 1952, p.62)	the ability to resolve mental rotation problems quickly (Lohman, 1979).	"the ability to imagine how a stimulus array will appear from another perspective" (Lohman, 1979, p.127).
	"the ability to mentally rotate, manipulate, and twist two- and three-dimensional stimulus objects." (McGee, 1979, p.896)	"the ability to visualize objects in space, when rotated." (Carroll, 1993, p.209)	"understanding and operating on the relationships between the positions of objects in space with respect to one's own position." (Clements & Battista, 1992, p.444)
	"comprehension and performance of imagined movements of objects in two- and three-dimensional space." (Clements & Battista, 1992, p.444)	"the ability to mentally rotate objects in two dimensions" (Contero et al., 2005, p.25)	"understanding and operating on relationships between different positions in space, at first with respect to one's own position and your movement through it, and eventually from a more abstract perspective that includes maps and coordinates at various scales." (Sarama & Clements, 2009, p.161)
	"the ability to understand accurately three-dimensional objects from their two-dimensional representation." (Barnea, 2000, p.308)	"the ability to visualise the effects of operations such as rotation, reflection and inversion, or to mentally manipulate objects." (Barnea, 2000, p.308)	"an ability to perceive spatial patterns or maintain orientation with respect to objects in space." (McGee, 1979, p. 892)
	"the mental manipulation and integration of stimuli consisting of more than one part or movable parts." (Olkun, 2003, p.2)	"the ability to comprehend the spatial configuration of objects or parts of an object and their relation to each other." (Maier, 1996, p.70)	"the ability to orient oneself physically or mentally in space" and it requires "a person's own orientation in any particular spatial situation." (Maier, 1996, p.71)
Tests	Paper Folding Test, Form Board Test, Surface Development Test (Ekstrom et al., 1976)	Flags Test (Thurstone & Thurstone, 1941)	Spatial Orientation Test (Guilford & Zimmerman, 1948)
	Purdue Spatial Visualization Test: Developments (Guay, 1977)	Purdue Spatial Visualization Test: Rotations (Guay, 1977)	Purdue Spatial Visualization Test: Views Test (Guay, 1977)
	Revised Minnesota Paper Form Board Test (Likert & Quasha, 1941)	Card Rotation Test and Cube Comparisons Tests (Ekstrom et al., 1976)	Middle Grades Mathematics Project (MGMP) Spatial Visualization Test (Winter et al., 1989)
	The Embedded Figures Test (Witkin, 1950)	Mental Rotation Tasks (Shepard & Metzler, 1971)	Spatial Orientation: Object Perspective/Map Perspective Tests (Kozhevnikov & Hegarty, 2001)

As an example of this discrepancy, while Olkun (2003, p. 2) defines spatial relations as “imagining the rotations of 2D and 3D objects as a whole body," Burnett and Lane (1980) and Olkun (2003) explain spatial visualization as a holistic and piece-by-piece imagination of the rotations of objects and their parts in 3D space. As can be understood from the definitions, while in the spatial relations ability, 2- and 3-dimensional objects are moved as a whole, in the spatial visualization subtest, the rotation of 3-dimensional objects happens with the whole and its parts. On the other hand, it has been frequently reported that while speed is more important in spatial relations tests, power is more important in spatial visualization test items (Olkun 2003; Pellegrino et al., 1984), problems in spatial relationships tests contain less complex stimuli than spatial visualization problems (Olkun, 2003), and more mental processing and coordination are required to solve spatial visualization problems (Pellegrino et al., 1984). In problems about spatial relations, the students have to find the rotated or twisted version of the original figure from among a group of objects given on a piece of paper (Olkun, 2003; Pellegrino et al., 1984). Pellegrino and Kail (1982) stated that spatial relations tests include problems measuring 2D and 3D mental rotation and cube comparison abilities. The tests measuring spatial visualization include form board problems (Linn & Petersen, 1985; Olkun, 2003; Pellegrino & Kail, 1982), paper folding (Contero et al., 2005; Linn & Petersen, 1985; McGee, 1979; NRC, 2006; Olkun, 2003; Pellegrino & Kail, 1982), and surface development (Contero et al., 2005; Linn & Petersen, 1985; Olkun, 2003; Pellegrino & Kail, 1982).

Lohman (1979) suggests that in a valid spatial orientation test, subjects must imagine being redirected in space and then interpret the situation. Spatial orientation tasks do not require moving an object mentally; only the perceptual perspective of the person viewing the object is changed or moved (Tartre, 1990). Measuring spatial orientation ability is difficult because it requires mental rotation of the stimulus rather than the rotation of the picture itself (Lohman, 1979). Tartre (1990) pointed out that there is no consensus among researchers on the classification of spatial orientation tasks, and stated that spatial ability tasks may involve organizing a visual representation, reorganizing, interpreting, seeing it, or seeing it from a different angle, but by moving the object mentally. Problems used to measure spatial orientation ability include finding directions on a map (Campos & Campos-Juanatey, 2020; Kozhevnikov & Hegarty, 2001), imagining the view of an object from different angles, determining the number of cubes in an object made up of cubes (Winter et al., 1989), finding the view of an object in a cube-shaped glass bell from different angles (Guay, 1977), etc.

1.2. Rationale

There are numerous spatial ability tests referred to in the relevant literature (see Table 2). As a result of the rapid proliferation of spatial ability tests, different researchers have given different names to similar factors or, conversely, the same names were used to describe different factors, which has measured the components of spatial ability even more complicated (Eliot & Hauptman, 1981). These confusions also affected the results of the factor analysis studies conducted to determine test structures. In a study, Carroll (1993) re-analyzed factor analytic studies in the literature and found that items are not always consistently loaded on relevant factors due to considerable confusion in the identification of factors. D'Oliveira (2004) also argued that variations in the format of the tests and specific administration procedures can be responsible for inconsistent results. Thus, D'Oliveira (2004) especially emphasized the importance of clarifying the spatial factor or capability covered by the test items, regardless of the names assigned to test items (or tasks) by previous researchers (Carroll, 1993). Due to this confusion in the literature, in this study, firstly, the definitions for the factors of spatial ability and question types were examined in detail and differentiated. Eventually, it was aimed to develop a spatial ability test in which question types and factors are clearly defined and statistically tested. The current test is planned to cover the three factors of spatial ability (spatial relations, spatial visualization, and spatial orientation) most commonly mentioned in the relevant literature. It is intended to expand the scope of the test by involving different types of problems in each factor.

Eliot and Hauptman (1981) pointed out that items can yield different factor loadings in different samples. This situation reveals the importance of developing the test in accordance with the characteristics of a particular group and testing its psychometric properties. What makes this test distinct from its antecedents is that while usually the same tests are used for different groups (Bakker, 2008; Battista et al., 1982; Ekstrom et al., 1976; Guay, 1977; Hegarty & Waller, 2004; Kozhevnikov & Hegarty, 2001; Lord, 1985; Sorby & Baartmans, 2000), the present test was developed specifically for university students studying at different programs/departments or candidate university students who plan to study programs that require spatial capability and to assess their professional competencies. In addition, several studies (e.g., Kim & Irizarry, 2021; Olkun et al., 2009; Patkin & Dayan, 2013; Sisman et al., 2021) have put forward the idea that spatial ability can be improved through well-designed training programs. In this context, it will be possible to accurately measure the spatial ability development among students and reveal the effect of the training only by using a valid and reliable spatial ability test.

On the other hand, the training applied may reveal different effects on the level of spatial ability in different cultures. For example, Turgut and Nagy-Kondor (2013) found a significant difference between the spatial visualization scores of Hungarian and Turkish pre-service mathematics teachers, favoring the former. Olkun et al. (2009) compared the initial spatial skills of primary school teacher candidates in four countries, i.e., Taiwan, Finland, the United States, and Türkiye, and evaluated the development of these skills through interactive computer programs. As a result, it was seen that the spatial visualization scores were the highest among Finnish students, followed by Taiwanese students, and the scores of the American and Turkish students were very close to each other. However, the researchers pointed out that while students from two eastern countries, Türkiye and Taiwan, made progress after the implementation, students from the USA and Finland did not make sufficient progress. Researchers stated that this situation may be due to cultural differences and suggested that the reason why spatial education is more successful in Taiwan and Türkiye than in the USA and Finland is that the former countries have relatively more formal class cultures. It is noteworthy that the spatial ability levels of Turkish students were reported as rather low in both of the abovementioned comparative studies. This reveals the importance of researching the spatial abilities of Turkish students. According to the literature review studies examining the tendency of spatial ability studies in Türkiye (Dokumacı Sütçü, 2021; İpekoğlu et al., 2020; Ozcakir Sumen, 2019), most of the studies were conducted with secondary school students and the effect of a particular teaching method (mostly computer-assisted teaching) on spatial ability was investigated. Since the transition to Piaget's formal operations stage coincides with the secondary school level, it is very important to focus on the development of students' spatial abilities during this period. However, since the spatial ability is very important in many professions, it is thought that the development of a spatial ability test to be used to measure the spatial ability levels of the students who are to get professional education at the university will be useful for researchers, educators, and curriculum developers.

The most commonly used tests were developed in the 1970s, and there are concerns about their psychometric properties since they have been administered to a wide range of different groups. In Türkiye, Purdue Spatial Visualization Test developed by Guay (1977) and the MGMP Spatial Ability Test developed by Michigan State University mathematics department faculty members (1983) were generally used to measure students' spatial visualization skills (İpekoğlu et al., 2020). However, the adaptation of the tests used in Turkish culture mainly concentrated on textual translation, and the equivalence of the tests in Turkish culture and their psychometric properties at the applied level were not adequately examined (Sevimli, 2009). Therefore, the present study is also promising because a more comprehensive test development procedure has been followed following a sequential exploratory mixed methods research design (qualitative followed by quantitative phases) and psychometric properties were tested through comprehensive analysis (item analysis, confirmatory factor analysis, reliability analysis, and the difference between 27% of the lower and upper groups).

2. METHOD

2.1. Design

The sequential exploratory mixed method was used to design this study, which aimed to develop a spatial ability test. The sequential exploratory mixed method is a common way of developing quantitative instruments, wherein in the first stage, the researcher starts to explore the subject using qualitative methods and then continues to validate the instrument using quantitative methods based on the themes from the first stage (Creswell & Plano Clark, 2011). In this study, qualitative methods (literature review, expert opinions, and student opinions) were used to develop the initial test form; and quantitative methods were used to test the psychometric properties regarding content validity, construct validity, and reliability.

2.2. Study Group

In the qualitative stage, 10 students (Female= 7, Male= 3) studying an elementary school mathematics education program were consulted to check the clarity, comprehensibility, and suitability of the draft test items. In addition, an expert panel consisting of 8 scholars (4 mathematics experts, 3 mathematics education experts, and one measurement and evaluation specialist) was consulted for their opinions about the content and face validity of the draft test.

In the quantitative stage, the validity and reliability studies of the test were conducted with a total of 456 university students (58% female), studying at different departments/programs of Malatya İnönü University, a state university located in eastern Türkiye. Participants were chosen from departments/programs where either recruitment or studying is considered to be facilitated by possessing good spatial abilities. Accordingly, participants involved 29 students (38% female) from the Graphic Design program, 61 students (7% female) from the Civil Engineering Department, 47 students (21% female) from the Mechanical Engineering Department, 266 students (77% female) from the Elementary School Mathematics Education Program, 39 students (67% female) from the Landscape Architecture Department, and 14 students (50% female) from the Art Teaching Program.

2.3. Procedure

In the development process of the spatial ability test, the stages of test development were followed, which included: 1) determining the purpose of the test; 2) determining the scope of the test, 3) determining test properties and writing items, 4) validity and reliability studies, and 5) preparing a guide for the test. Accordingly, first, an overall plan regarding the test development process was prepared by the researchers, which was then evaluated by a measurement and evaluation specialist. The plan was revised in accordance with the experts' opinions and put into practice as described below:

2.3.1. Determining the purpose of the test

The purpose of the spatial ability test is to measure the spatial ability levels of university students in a valid and reliable way.

2.3.2. Determining the scope of the test

Downing (2006) emphasized that determining the content of the test is one of the most important tasks at the earliest stages of the test development process. Due to the critical importance of the scope of the spatial ability test, we set out with a detailed literature review first. As a result of the comprehensive literature review, it was seen that spatial ability has a multifactorial structure, and different researchers explain spatial ability under different factors (D'Oliveira, 2004; Lohman, 1993; Mohler, 2008). Since it would not be possible to include all factors of spatial ability mentioned in the studies in terms of usefulness, reliability, and content validity, it was decided to include three domains of spatial ability most commonly referred to in the literature: *spatial relations, spatial visualization, and spatial orientation.* After deciding on the factors to be included in the test, a second literature review was conducted to examine the conceptual and operational (how they are measured) definitions of these factors. The definitions of the factors of the spatial ability test developed in this study are presented in Table 3.

Table 3. *The definitions of the factors of spatial ability test.*

Factor	Definition
Spatial Ability	the ability to generate, retain, retrieve, manipulate, interpret, reorganize the mental representations of visual objects by perceiving their forms and positions (Carroll, 1993; Linn & Petersen, 1985; Lohman, 1993; NRC, 2006; Tartre, 1990).
Spatial Relations	the ability to mentally manipulate 2D and 3D objects as a whole with processes such as rotation, reflection, and inversion (Barnea, 2000; Carroll, 1993; Contero et al., 2005; Olkun, 2003).
Spatial Visualization	the ability to mentally rotate, manipulate, and twist a 3-dimensional object composed of more than one part or movable parts in a holistic and piece-by-piece (Burnett & Lane, 1980; McGee, 1979; Olkun, 2003).
Spatial Orientation	the ability to understand the relations between the positions of objects in space relative to one's own position (Clements & Battista, 1992; Sarama & Clements, 2009) and to imagine how an object will look in space from a different perspective by mentally orienting oneself (Barnea, 2000; Contero et al., 2005; Lohman, 1979; Maier, 1996; McGee, 1979).

2.3.3. Determining test properties and writing items

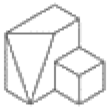
In this study, spatial ability test items were planned to be developed in a multiple-choice format with 5 options. A total of 38 original test items in different problem types were developed by the researchers. In addition, 2 items about rotating 3D objects (Item 12, Item 13) in the draft spatial relations subtest and 3 items unfolding 3D objects (Item 27, Item 28, Item 29) in the draft spatial visualization subtest were driven from Guay's (1977) Purdue Spatial Visualization Test: Rotations and Purdue Spatial Visualization Test: Developments tests, respectively. As a result, an item pool of 43 items was developed, including 17 items in the spatial relations subtest, 12 items in the spatial visualization subtest, and 14 items in the spatial orientation subtest as can be seen in Table 4.

Table 4. *Components, categories, and numbers of items.*

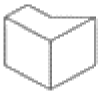
Component	Category	Item No	Number of items
Spatial Relations	Card rotation	1, 2, 3	3
	Rotating the 2D figures and symmetry	4, 5, 6, 7, 8	5
	Rotating the 3D figures	9, 10, 11, 12, 13	5
	Comparing cubes	14, 15, 16, 17	4
Spatial Visualization	Unfolding cubes	18, 19, 20, 21	4
	Cutting paper	22, 23, 24	3
	Unfolding 3D objects	25, 26, 27, 28, 29	5
Spatial Orientation	The view of an object made up of cubes from different angles	30, 31, 32, 33, 37, 38, 39, 40	8
	Number of cubes	34, 35, 36	3
	The view of an object in a cube-shaped glass bell from different angles	41, 42, 43	3

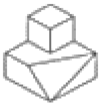
Examples of items are presented below.

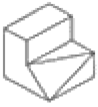
Spatial Relation:

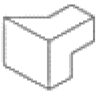


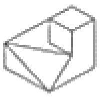
10. Which of the following is the rotated form of the object on the left?

A)


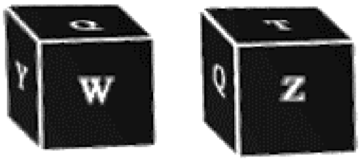
B)


C)


D)


E)


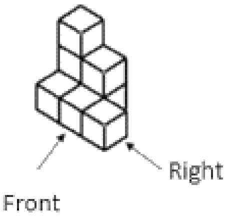
Spatial Visualization:



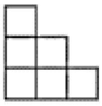
20. Images of the same cube in different positions are presented on the left labeled with letters X, Y, T, Z, Q, and W. Which letter is located opposite the letter X?

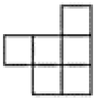
A) Y B) Q C)W D) Z E) T

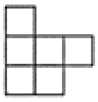
Spatial Orientation:

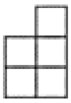


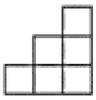
30. On the left is the FRONT-RIGHT corner view of an image. Which of the following is the view of the same object from the REAR?

A)


B)


C)


D)


E)


2.3.4. Validity and reliability studies

An expert panel consisting of 4 mathematics experts, 3 mathematics education experts, and 1 measurement and evaluation specialist was asked to evaluate the content and face validity of the test. The evaluation criteria included scientific accuracy, comprehensibility and responsiveness of the question roots and options, and the suitability of the figures. The experts evaluated each item using the 4-point scale offered by Davis (1992): 4-Highly relevant, 3-Quite relevant, 2-Somewhat relevant, and 1-Not relevant. The criteria for the Content Validity Index (CVI), which is computed as the number of experts rating an item either 3 or 4, divided by the total number of experts, is set to a minimum of 0.80 (Davis, 1992). Based on the expert ratings, CVIs for all items were found to satisfy the minimum criteria of 0.80. In addition, the revision suggestions from the experts were done and the draft test form was developed with 43 items. Further, to assess the clarity, understandability, and appropriateness of the test form to the target audience, within the scope of the think-aloud protocol, another 10 prospective primary school mathematics teachers were asked to take the test and verbally express their mental processes while solving each test item (Irwing et al., 2018). This way the test items were checked to ensure whether they can measure the constructs which they were actually meant to test. To ensure the reliability and validity of the test results, the figures and the question roots were checked for readability during the preparation and printing of the booklets (Downing, 2006).

Next, the test was applied to 456 university students to examine the item and test statistics. To test the construct validity of the instrument, item difficulty index, item discrimination index, and item-total correlation coefficients were calculated, and the significance of *t* values regarding the differences between 27% lower and upper groups were examined. In addition, a second-order confirmatory factor analysis was performed to test the 3- factor (spatial relations, spatial visualization, and spatial orientation) construct of the spatial ability test. The reliability of the scores obtained from the test was calculated using KR-20 and Split-Half (odd-even) with Spearman-Brown reliability coefficients.

2.3.5. Preparation of a guide for test users

It is planned to provide users with information about the application of the test through a guide, which specifies the purpose of the test, its theoretical background, scoring procedures, and descriptive statistics at the end of the study.

2.4. Data Analysis

The analysis of the data obtained from 456 participants was made via the Test Analysis Program (TAP) (Brooks & Johanson, 2003), SPSS 22, and Lisrel software programs. Correct and incorrect or blank answers were scored 1-0, respectively. The skewness and kurtosis coefficients for the data set were estimated at 0.363 and 0.322, respectively. Since the skewness and kurtosis coefficients were within the acceptable range, it was understood that the data set comes from a normal distribution. While item difficulty refers to the percentage or probability (P) of test takers who answer the item correctly (Ebel & Frisbie, 1991; Hingorjo & Jaleel, 2012; Wendler & Walker, 2006), item discrimination is the tendency of an item to be answered correctly by test takers who are strong in terms of the skill or knowledge intended to be measured and to be answered incorrectly by test takers who are not strong in this respect (Livingston, 2006). The item difficulty indices were kept in the 0.30 to 0.70 range, with fewer items in the easier or more difficult ranges, because in large-scale standardized tests, test taker levels are typically assumed to be normally distributed, and items in the middle range of difficulty have the most variance and the greatest potential to discriminate test takers (Bandalos, 2018, p. 122). Hambleton and Jirka (2006) categorized values around 0.25 as "difficult," values around 0.50 as "moderate," and values around 0.75 as “easy” in terms of item difficulty. Items with a discrimination index of 0.40 or higher were considered very good; items with a discrimination index of 0.30 to 0.39 were considered reasonably good but could be developed; items with a discrimination index of 0.20 to 0.29 were considered poorly discriminative but could be corrected or improved; and items with a discrimination index of 0.19 or lower were considered very poor and could not be corrected or improved. The ideal item-total correlation coefficient was set to a minimum of 0.30 (Wendler & Walker, 2006). In addition, to estimate how discriminative the individual test items are, the differences between the scores of the 27% upper and lower groups were compared using independent samples t-test since the scores were close to the normal distribution (skewness and kurtosis values ± 2 (Cameron, 2004)). The significance level was set to $p < 0.05/27 = 0.002$ (n= 27 t-tests for differences between the scores of the 27% lower and upper groups) with a Bonferroni correction (Abdi, 2010).

3. FINDINGS

3.1. Findings about the Item Analysis of the Spatial Ability Test

The construct validity of the Spatial Ability Test was tested through item analysis. Accordingly, item difficulty indices, item discrimination indices, and item-total correlation coefficients calculated for the preliminary spatial ability test consisting of 43 items are presented in Table 5.

Table 5. *Results of item analysis.*

Item no	Item difficulty(P)	Item discrimination (d)	Item-total correlation (r)
1	0.68	0.39	0.36
2	0.73	0.47	0.43
3	0.38	0.34	0.32
4	0.68	0.54	0.45
5	0.68	0.41	0.36
6	0.50	0.28	0.27
7	0.44	0.37	0.37
8	0.57	0.46	0.37
9	0.51	0.35	0.32
10	0.72	0.46	0.43
11	0.48	0.36	0.33
12	0.35	0.35	0.32
13	0.38	0.30	0.26
14	0.40	0.31	0.27
15	0.38	0.21	0.19
16	0.61	0.42	0.37
17	0.54	0.56	0.44
18	0.45	0.32	0.31
19	0.23	0.06	0.07
20	0.51	0.52	0.43
21	0.20	0.12	0.14
22	0.41	0.33	0.32
23	0.30	0.38	0.41
24	0.20	0.21	0.24
25	0.24	0.21	0.24
26	0.29	0.13	0.07
27	0.31	0.15	0.14

Item no	Item difficulty(P)	Item discrimination (d)	Item-total correlation (r)
28	0.16	0.20	0.29
29	0.20	0.24	0.30
30	0.48	0.43	0.35
31	0.19	0.16	0.18
32	0.18	0.05	0.17
33	0.72	0.49	0.43
34	0.37	0.42	0.37
35	0.34	0.34	0.29
36	0.31	0.15	0.12
37	0.26	0.19	0.17
38	0.31	0.27	0.25
39	0.29	0.38	0.34
40	0.29	0.33	0.31
41	0.47	0.47	0.36
42	0.44	0.34	0.33
43	0.38	0.45	0.38

In Table 5, it was decided to exclude 19, 21, 26, 27, 31, 32, 36, and 37 items with a discrimination index below 0.20. To decide whether items with an item discrimination index between 0.20-0.40 to be corrected or excluded from the test, their item-total correlations were examined. Accordingly, it was decided to exclude items 6, 13, 14, 15, 24, 25, 35, and 38 with item-total correlations below 0.30. Table 4 presents components, categories, and numbers of items. During the item analysis, attention was paid to retain at least 2 items in each category in the final test so that the content validity was not impaired. Thus, despite their relatively low item discrimination indices, items 28 (d= 0.20) and 29 (d= 0.24) were decided to be kept in the test to ensure content validity. Starting from the item with the lowest discrimination index and item-total correlation value, the problematic items were removed successively and the item analysis was repeated. Item analysis results for the final test are presented in Table 6.

According to Table 6, the difficulty indices of the items in the final test ranged between 0.16 and 0.73 (mean = 0.475). Accordingly, 4 items in the final test were difficult (items 28, 29, 39, and 40), 3 items were easy (items 2, 10, and 33), and the remaining 20 items were moderate in terms of difficulty. Discrimination indices of the items ranged between 0.23 and 0.55 (mean = 0.421), and the item-total correlation coefficient ranged between 0.28 and 0.47 (mean = 0.381). In the final test, items 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 16, and 17 measure spatial relations ability, 18, 20, 22, 23, 28, and 29 items measure spatial visualization ability, and items 30, 33, 34, 39, 40, 41, 42, 43 measure spatial orientation ability. The average difficulty index of the 27 items in the Spatial Ability test was 0.475; while the average discrimination index was 0.421 and the average item-total correlation was 0.381.

Table 6. *Results of item analysis and descriptive analysis for the items in the final test.*

Item no	Item difficulty (P)	Item discrimination (d)	Item-total correlation (r)
1	0.68	0.43	0.38
2	0.73	0.49	0.47
3	0.38	0.42	0.37
4	0.68	0.51	0.44
5	0.68	0.42	0.37
7	0.44	0.42	0.37
8	0.57	0.48	0.41
9	0.51	0.38	0.34
10	0.72	0.45	0.45
11	0.48	0.43	0.36
12	0.35	0.37	0.34
16	0.61	0.44	0.38
17	0.54	0.53	0.45
18	0.45	0.38	0.35
20	0.51	0.55	0.44
22	0.41	0.33	0.33
23	0.30	0.45	0.41
28	0.16	0.27	0.30
29	0.20	0.23	0.28
30	0.48	0.43	0.38
33	0.72	0.47	0.43
34	0.37	0.43	0.40
39	0.29	0.33	0.34
40	0.29	0.36	0.36
41	0.47	0.43	0.36
42	0.44	0.44	0.37
43	0.38	0.50	0.41

3.2. Differences between 27% lower and upper group scores

Another method used to test the construct validity of the test through the discrimination potential of the items is to compare, for each item, the average scores from 27% lower and upper groups using the independent t-test. The results of the independent samples t-test regarding the comparison of the averages of the 27% lower group (n = 144) and 27% upper group (n = 131) are presented in Table 7.

When Table 7 is examined, statistically significant differences were found between the lower and upper groups of 27% for all items (p <0.002). Therefore, in addition to item discrimination and item-total correlation coefficient analyses, it was proven once again that each item is able to significantly distinguish between the upper group with the highest spatial ability and the lower group with the lowest spatial ability.

Table 7. *t-test results regarding the significance of the differences between the scores of the lower and upper groups (27%).*

Item no	Group	Mean	Std. Deviation	*t*	*df*	*p*
Item 1	Lower 27%	0.46	0.50	-8.516	245.781	.000*
	Upper 27%	0.89	0.32			
Item 2	Lower 27%	0.46	0.50	-10.591	203.058	.000*
	Upper 27%	0.95	0.23			
Item 3	Lower 27%	0.16	0.37	-7.928	238.462	.000*
	Upper 27%	0.58	0.50			
Item 4	Lower 27%	0.42	0.50	-10.707	223.761	.000*
	Upper 27%	0.92	0.27			
Item 5	Lower 27%	0.47	0.50	-8.457	241.682	.000*
	Upper 27%	0.89	0.31			
Item 7	Lower 27%	0.19	0.39	-7.708	248.363	.000*
	Upper 27%	0.60	0.49			
Item 8	Lower 27%	0.30	0.46	-9.087	272.999	.000*
	Upper 27%	0.78	0.42			
Item 9	Lower 27%	0.32	0.47	-6.845	271.439	.000*
	Upper 27%	0.70	0.46			
Item 10	Lower 27%	0.47	0.50	-9.444	222.183	.000*
	Upper 27%	0.92	0.27			
Item 11	Lower 27%	0.26	0.44	-7.894	273	.000*
	Upper 27%	0.69	0.46			
Item 12	Lower 27%	0.16	0.37	-7.014	236.955	.000*
	Upper 27%	0.53	0.50			
Item 16	Lower 27%	0.40	0.49	-8.530	263.698	.000*
	Upper 27%	0.84	0.37			
Item 17	Lower 27%	0.26	0.44	-10.295	273	.000*
	Upper 27%	0.79	0.41			
Item 18	Lower 27%	0.27	0.45	-6.753	265.644	.000*
	Upper 27%	0.65	0.48			
Item 20	Lower 27%	0.26	0.44	-10.806	272.897	.000*
	Upper 27%	0.81	0.39			
Item 22	Lower 27%	0.29	0.46	-5.721	266.061	.000*
	Upper 27%	0.62	0.49			
Item 23	Lower 27%	0.12	0.32	-8.894	220.074	.000*
	Upper 27%	0.57	0.50			
Item 28	Lower 27%	0.04	0.20	-6.171	173.113	.000*
	Upper 27%	0.31	0.47			
Item 29	Lower 27%	0.11	0.32	-4.719	222.117	.000*
	Upper 27%	0.34	0.48			
Item 30	Lower 27%	0.27	0.45	-7.902	273	.000*
	Upper 27%	0.70	0.46			
Item 33	Lower 27%	0.47	0.50	-10.142	209.860	.000*
	Upper 27%	0.94	0.24			
Item 34	Lower 27%	0.19	0.39	-8.027	249.299	.000*

Item no	Group	Mean	Std. Deviation	*t*	*df*	*p*
	Upper 27%	0.62	0.49			
Item 39	Lower 27%	0.14	0.35	-6.231	228.730	.000*
	Upper 27%	0.47	0.50			
Item 40	Lower 27%	0.14	0.35	-6.947	228.396	.000*
	Upper 27%	0.50	0.50			
Item 41	Lower 27%	0.28	0.45	-7.912	273	.000*
	Upper 27%	0.71	0.46			
Item 42	Lower 27%	0.24	0.43	-8.180	263.595	.000*
	Upper 27%	0.68	0.47			
Item 43	Lower 27%	0.15	0.36	-8.180	263.595	.000*
	Upper 27%	0.66	0.48			

*p<0.002 (with a Bonferroni correction of 0.05/27=0.002)

3.3. Second-order Confirmatory Factor Analysis

Second-order confirmatory factor analysis (CFA) was performed to test the three-factor construct (spatial visualization, spatial relations, and spatial orientation) of the 27-item spatial ability test. The model was estimated using the asymptotic covariance matrix and analyzed using the Diagonally Weighted Least Squares method in the Lisrel software program (Jöreskog & Sörbom, 1993; Kline, 2011). While evaluating CFA results, the goodness of fit indices was considered excellent when $\chi^2/df \leq 2$; GFI, AGFI, CFI, IFI, NNFI $\geq$ 0.95; RMSEA, SRMR $\leq$ 0.05, and they were indicated acceptable when $\chi^2/df \leq 5$; GFI, AGFI, CFI, IFI, NNFI $\geq$ 0.90; SRMR, RMSEA $\leq$ 0.08 (e.g. Brown, 2006; Hair et al., 2014; Hu & Bentler, 1999; Jöreskog & Sörbom, 1996; Tabachnick & Fidell, 2013). As a result of the first analysis, the goodness of fit indices was estimated χ^2/df = 1.39 (445.99/321), RMSEA =0.029, SRMR = 0.075, GFI = 0.96, AGFI = 0.95, CFI = 0.98, IFI = 0.98, NNFI = 0.98. According to these values, χ^2/df, RMSEA, GFI, AGFI, CFI, IFI NNFI values were excellent, SRMR value indicated acceptable. Standardized factor loadings ranged between 0.60 and 0.86 in the spatial relations subtest, between 0.69 and 0.88 in the spatial visualization subtest, and between 0.59 and 0.82 in the spatial orientation subtest. These results suggested that the three-factor construct of the spatial ability test is confirmed and a total Spatial Ability score can be calculated for all 27 items. The path diagram is presented in Figure 1.

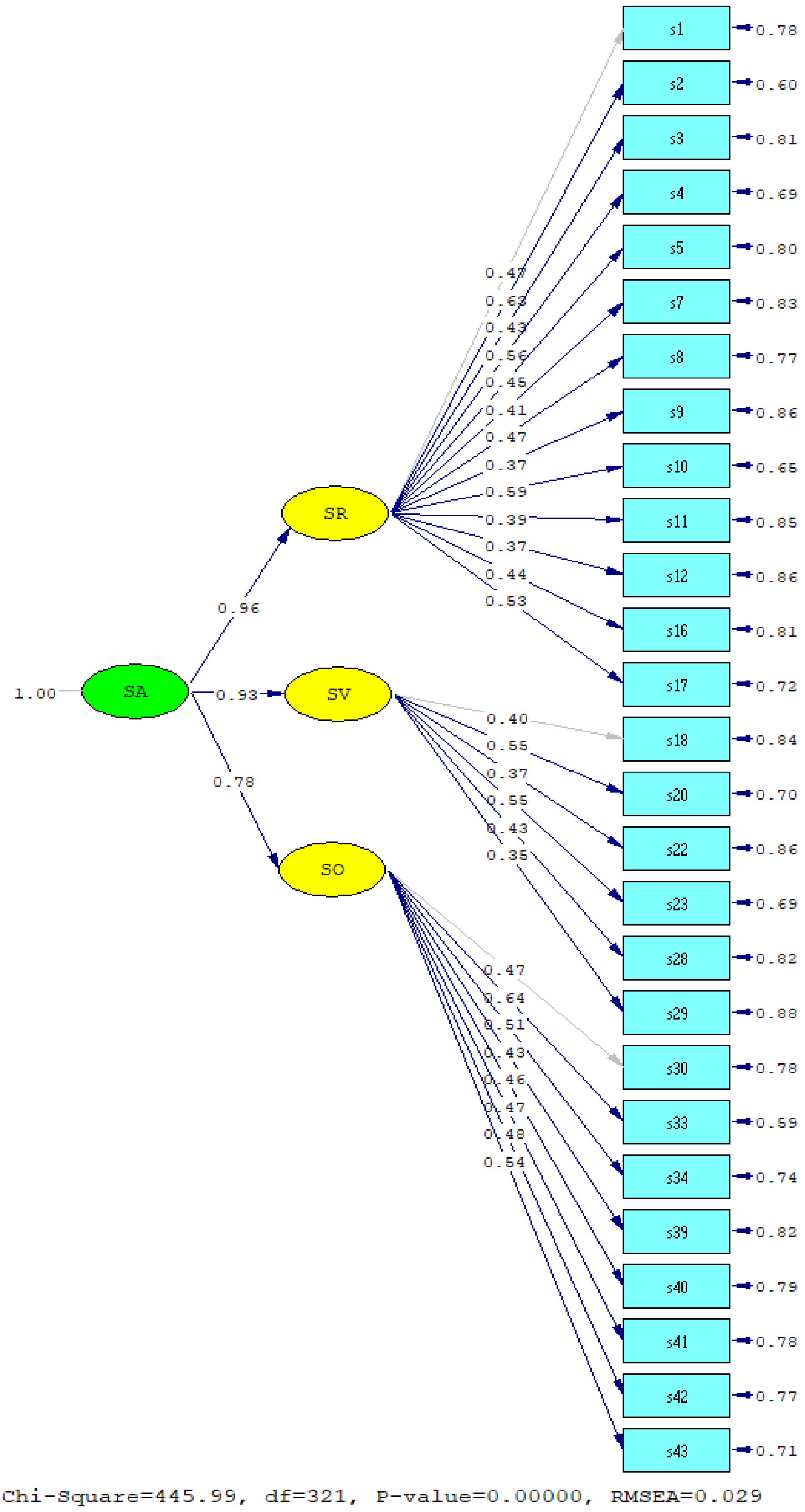
3.4. Reliability Analysis

The KR-20 reliability coefficient of the total test was estimated 0.775, and Split-Half (odd-even) with Spearman-Brown was estimated 0.798 suggesting acceptable internal consistency. Wells and Wollack (2003) put that the minimum value of the reliability coefficient is expected to be 0.70. Based on this reference value, it can be said that the reliability coefficients of the test are sufficient for the whole test.

3.5. The Guide for Test Users

The Spatial Ability Test is a multiple-choice test to measure university students' spatial abilities. The final test has three sub-tests with 27 items each offering 5 options: 13 items in the spatial relations subtest (items 1-13), 6 items in the spatial visualization subtest (items 14-19), and 8 items in the spatial orientation subtest (items 20-27). Correct and incorrect/blank answers are scored 1-0 respectively. For the test, a student has a minimum possible score of 0 and a maximum possible score of 27. The high scores are indicative of good level spatial ability, whilst low scores are indicative of low spatial ability. For average scores, 0-9.00 points can be interpreted as low, 9.01-18.00 points as medium, and 18.01-27.00 points as good spatial ability skills. The mean spatial ability score for the participants of this study was 12.85 (s= 4.89), indicating medium level of spatial ability.

Figure 1. *Path diagram of the model.*



Note. SA: Spatial Ability; SR: Spatial Relations; SV: Spatial Visualization; SO: Spatial Orientation

4. DISCUSSION and CONCLUSION

This study aimed to develop a useful test to measure university students' spatial abilities in a valid and reliable way. To develop a comprehensive and focused instrument, it was planned to develop items related to spatial relations, spatial visualization, and spatial orientation abilities as the subtests of spatial ability. The validity studies of the spatial ability test were carried out in detail and meticulously, and evidence for three criteria was collected to determine the validity of the test: content validity, face validity, and construct validity.

Prior to the quantitative pilot study, the opinions of experts evaluated according to the Davis (1992) technique proved the adequate level of content and face validity. In the process of testing the construct validity, item analysis was performed first. Crocker and Algina (2008) suggested that when developing a test, it was aimed to produce a final test including an optimum number of items, which meet the required reliability and validity criteria. Therefore, taking into account the usefulness of the test, it was aimed to develop a valid and reliable test with a minimum number of items while preserving the content validity. Accordingly, during item analysis, items with poor discrimination indexes and item-total correlation coefficients were successively excluded from the test, and a final test form with 27 items was obtained. The final test included 13 items in the spatial relations ability subtest, 6 items in the spatial visualization ability subtest, and 8 items in the spatial orientation ability subtest.

When the average discrimination indices for the spatial ability test and its subtests are examined, it can be said that the test as a whole and its subtests are highly discriminative (Ebel & Frisbie, 1991; Wells & Wollack, 2003). The average item-total correlation coefficients for the spatial ability test and its subtests indicated the adequacy of discrimination and internal consistency (Wendler & Walker, 2006). Moreover, we found statistically significant differences between the 27% lower and upper group scores for each item, which provided additional evidence for the existence of the items' discrimination, as the difference between the upper and lower groups of 27% reveals a more sensitive and stable item discrimination index about the test items (Crocker & Algina, 2008; Diederich, 1973).

The average difficulty index of the test is moderate (Hambleton & Jirka, 2006). Hingorjo and Jaleel (2012) point out that item difficulty and item discrimination indexes are generally interrelated. Similarly, it is well known that test developers should avoid including items that are answered correctly or incorrectly by the majority of students since such items would have standard deviations close to zero and cannot distinguish students with different ability levels (Crocker & Algina, 2008; Wells & Wollack, 2003; Wendler & Walker, 2006). Since the variance would be maximum when the item difficulty is 0.50, it has been suggested that most of the test items should be a moderate difficulty (around 0.50) to discriminate well between people with a wide range of abilities (Crocker & Algina, 2008; Gronlund, 1977; Wendler & Walker, 2006). The average difficulty level of the items in the final spatial ability test developed in this study was also around 0.50. Accordingly, it can be said that the average difficulty of the test increases the variance and contributes to the potential of the test to distinguish individuals with high and low spatial abilities. When the difficulty of the subtests is examined, it can be said that the average difficulty values of the spatial relations and spatial orientation tests are closer to moderate difficulty; however, the average difficulty of the spatial visualization test indicated a rather difficult test. Several studies (Linn & Petersen, 1985; Lohman, 1979; Olkun, 2003; Pellegrino et al., 1984) report that spatial visualization is more complex than other subskills. The result obtained in this study regarding the difficulty of the spatial visualization test compared to other subtests is in line with the literature. Thus, in this study, it can be said that the construct validity of the test was provided according to the results obtained from the item analysis of the test.

As a part of construct validity studies, the three-factor construct (spatial relations, spatial visualization, and spatial orientation) of the 27-item final test was examined with a second-order CFA. According to the widely accepted goodness of fit criteria (e.g., Brown, 2006; Hair et al., 2014; Tabachnick & Fidell, 2013), the goodness of fit indices for the three-factor construct were determined to be perfect, except for SRMR, which is also acceptable. Also, the standardized factor loadings were estimated between 0.59 and 0.88 and significant for all items. Brown (2006) stated that factor loadings greater than or equal to 0.30 or 0.40 are regarded as salient. The result of reliability analysis through KR-20 and Split-Half (odd-even) with Spearman-Brown coefficients proved to be favorable (Wells & Wollack, 2003). As a result, it can be said that the 27-item test is useful, valid, and reliable for measuring the spatial abilities of university students.

5. LIMITATIONS and FUTURE DIRECTIONS

The participants of this research are restricted to 456 students studying at the departments/programs of graphic design, civil engineering, mechanical engineering, elementary mathematics education, landscape architecture, and art education at a state university in eastern Türkiye. Therefore, the results of the research may not be practically generalized to students studying in all departments of the university and different levels (e.g, high school, graduate). In addition, the study may have shown context-based results and may limit the generalization of the results to other regions. Thus, the psychometric properties can be tested further with university students studying in different departments, with different levels (e.g, high school, graduate), or different regions of Türkiye. In addition, the participants in this study were recruited using the purposive and convenience sampling method. The psychometric properties of the test can be tested on a group determined by random assignment.

Considering the abovementioned limitations, the spatial ability test developed here can be used by high school guidance services to measure students' or candidates' spatial abilities to predict their potential for tertiary programs requiring such abilities as dentistry, medicine, architecture, engineering, navigation, mathematics, art, graphic design, etc. The same is also true for any admission committees that plan to measure candidates' spatial abilities for employment or program admission purposes. Moreover, the spatial ability test developed in this study can be used as a pre-posttest to test the effect of the potential intervention programs aiming at improving learners' spatial ability.

Acknowledgments

The findings of the research were presented as an oral paper at International EJER2020 Congress, held on 10-13 September at Eskişehir Anadolu University, Turkiye.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. Scientific and ethical principles were complied. It has been confirmed by İnönü University Social and Human Sciences Scientific Research Ethics Committee (24/03/2022- Ethics Committee Number: 2022/6-24). The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Kubra Acikgul: Literature review, conceptualization, preparation of the item pool, data collecting, validity and reliability studies, writing- original draft preparation. **Suleyman Nihat Sad:** Literature review, conceptualization, validity and reliability studies, writing- original draft preparation, reviewing and editing. **Bilal Altay:** Preparation of the item pool, data collecting.

Orcid

Kubra Acikgul https://orcid.org/0000-0003-2656-8916
Suleyman Nihat Sad https://orcid.org/0000-0002-3169-2375
Bilal Altay https://orcid.org/0000-0002-2400-7122

REFERENCES

Abdi, H. (2010). Holm's sequential Bonferroni procedure. In N. Salkind (Eds.), *Encyclopedia of research design* (pp. 1-8). Sage Publication.

Bakker, M. (2008). *Spatial ability in primary school: Effects of the Tridio® learning material* [Master's thesis]. University of Twente.

Bandalos, D.L. (2018). *Measurement theory and applications for the social sciences* (1st ed.). Guilford Publications.

Battista, M.T., Wheatley, G.H., & Talsma, G. (1982). The importance of spatial visualization and cognitive development for geometry learning in preservice elementary teachers. *Journal for Research in Mathematics Education, 13*(5), 332-340. https://doi.org/10.5951/jresematheduc.13.5.0332

Barnea, N. (2000). Teaching and Learning about Chemistry and Modelling with a Computer managed Modelling System. In Gilbert J.K., Boulter C.J. (Eds.), *Developing Models in Science Education* (pp. 307-323). Springer. https://doi.org/10.1007/978-94-010-0876-1_16

Brooks, G.P., & Johanson, G.A. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement, 27*(4), 303-304. https://doi.org/10.1177/0146621603027004007

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (1st ed.). The Guilford Press.

Burnett, S.A., & Lane, D.M. (1980). Effects of academic instruction on spatial visualization. *Intelligence*, *4*(3), 233-242. https://doi.org/10.1016/0160-2896(80)90021-5

Cameron, A. (2004). Kurtosis. In M. Lewis-Beck, A. Bryman and T. Liao (Eds.). *Encyclopedia of social science research methods* (pp. 544-545). SAGE Publications, Inc.

Campos, A., & Campos-Juanatey, D. (2020). Measure of spatial orientation ability. *Imagination, Cognition and Personality*, *39*(4), 348-357. https://doi.org/10.1177/0276236619896268

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies* (No. 1). Cambridge University Press.

Clements, D.H. (1998). *Geometric and spatial thinking in young children*. National Science Foundation.

Clements, D.H., & Battista, M.T. (1992). Geometry and spatial reasoning. In D.A. Grouws (Eds.), *Handbook of research on mathematics teaching and learning* (pp. 420-464). Macmillan.

Contero, M., Naya, F., Company, P., Saorín, J.L., & Conesa, J. (2005). Improving visualization skills in engineering education. *IEEE Computer Graphics and Applications*, *25*(5), 24-31. https://doi.org/10.1109/MCG.2005.107

Contreras, M.J., Escrig, R., Prieto, G., & Elosúa, M.R. (2018). Spatial visualization ability improves with and without studying Technical Drawing. *Cognitive processing*, *19*(3), 387-397. https://doi.org/10.1007/s10339-018-0859-4

Creswell, J.W., & Plano Clark, V.L. (2011) *Designing and conducting mixed methods research* (2nd ed.). Sage Publications.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

D'Oliveira, T.C. (2004). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *The International Journal of Aviation Psychology*, *14*(1), 19-38. https://doi.org/10.1207/s15327108ijap1401_2

Davis, L.L. (1992). Instrument review: getting the most from a panel of experts. *Applied Nursing Research*, 5, 194-197. https://doi.org/10.1016/S0897-1897(05)80008-4

Diederich, P.B. (1973). Short-cut statistics for teacher-made tests. Educational testing service. https://files.eric.ed.gov/fulltext/ED081785.pdf

Dokumacı Sütçü, N. (2021). Research tendencies towards spatial ability in Turkey. *International Journal* of Society *Researches*, *17*(36), 2605-2636. https://doi.org/10.26466/opus.839496

Downing, S.M. (2006). Twelve steps for effective test development. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Lawrence Erlbaum Associates.

Ebel, R.L., & Frisbie, D.A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall.

Ekstrom, R.B., French, J.W., Harman, H.H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Educational Testing Service.

Eliot, J., & Hauptman, A. (1981). Different dimensions of spatial ability. *Studies in Science Education*, *8*(1), 45-66. https://doi.org/10.1080/03057268108559886

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Gilligan, K.A., Flouri, E., & Farran, E.K. (2017). The contribution of spatial ability to mathematics achievement in middle childhood. *Journal of experimental child psychology*, *163*, 107-125. https://doi.org/10.1016/j.jecp.2017.04.016

Gronlund, N.E. (1977). *Constructing achievement tests* (2nd ed.). Prentice-Hall.

Guay, R.B. (1977). *Purdue spatial visualization test.* Purdue Research Foundation.

Guilford, J.P., Fruchter, B., & Zimmerman, W.S. (1952). Factor analysis of the army air forces sheppard field battery of experimental aptitude tests. *Psychometrika, 17*(1), 45–68. https://doi.org/10.1007/BF02288795

Guilford, J.P., & Zimmerman, W.S. (1948). The Guilford-Zimmerman Aptitude Survey. *Journal of applied Psychology*, *32*(1), 24-34. https://doi.org/10.1037/h0063610

Hair, J.F., Jr., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2014). *Multivariate data analysis* (7th ed.). Pearson New International Edition.

Hambleton, R.K., & Jirka, S.J. (2006). Anchor-based methods for judgmentally estimating item statistics. In Downing, S.M., & Haladyna, T.M. (Eds.). *Handbook of test development* (pp. 399-420). Lawrence Erlbaum Associates Publishers.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32*(2), 175-191. https://doi.org/10.1016/j.intell.2003.12.001

Hingorjo, M.R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association*, *62*(2), 142-147. https://www.jpma.org.pk/PdfDownload/3255

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

Irwing, P., Booth, T., & Hughes, D.J. (2018). *The Wiley handbook of psychometric testing. A Multidisciplinary Reference on Survey, Scale and Test Development.* Wiley-Blackwell.

İpekoğlu, A., Kepceoğlu, İ. & Biber, A.Ç. (2020). Thematic and methodological trends of graduate theses related to spatial ability: the case of Turkey. *International Journal of Contemporary Educational Studies*, *6*(2), 681-699. https://dergipark.org.tr/en/pub/intjces/issue/59193/829670

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8: Structural Equations Modeling with the SIMPLES Command Language*. Scientific Software, International.

Jöreskog, K.G. & Sörbom, D. (1996). *Lisrell 8 reference guide.* Scientific Software International.

Kell, H.J., Lubinski, D., Benbow, C.P., & Steiger, J.H. (2013). Creativity and technical innovation: Spatial ability's unique role. *Psychological Science*, *24*(9), 1831-1836. https://doi.org/10.1177/0956797613478615

Kerkman, D.D., Wise, J.C., & Harwood, E.A. (2000). Impossible "mental rotation" problems: A mismeasure of women's spatial abilities?. *Learning and Individual Differences*, *12*(3), 253-269. https://doi.org/10.1016/S1041-6080(01)00039-5

Kim, J., & Irizarry, J. (2021). Evaluating the use of augmented reality technology to improve construction management student's spatial skills. *International Journal of Construction Education and Research, 17*(2), 99-116. https://doi.org/10.1080/15578771.2020.1717680

Kline, R.B. (2011). *Principles and practice of structural equation modeling*. Guilford Press.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, *29*(5), 745-756. https://doi.org/10.3758/BF03200477

Likert, R., & Quasha, W.H. (1941). *Revised Minnesota Paper Form Board*. Psychological Corporation.

Linn, M., & Petersen, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, *56*(6), 1479-1498. https://www.jstor.org/stable/1130467

Livingston, S.A. (2006). Item Analysis. In Downing, S. M., & Haladyna, T. M. (Eds.). *Handbook of test development* (pp. 421-441). Lawrence Erlbaum Associates Publishers.

Lohman, D.F. (1979). Spatial ability: A review and reanalysis of the correlational literature. *Technical Report. No. 8*, Stanford University. https://apps.dtic.mil/sti/pdfs/ADA075972.pdf

Lohman, D.F. (1993, July 8). *Spatial ability and g*. Paper presented at the first Spearman Seminar, University of Plymouth, England. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7385&rep=rep1&type=pdf

Lord, T.R. (1985). Enhancing the visuo-spatial aptitude of students. *Journal of research in science teaching*, *22*(5), 395-405. https://doi.org/10.1002/tea.3660220503

Maier, P.H. (1996, March). Spatial geometry and spatial ability–How to make solid geometry solid. In *Selected papers from the Annual Conference of Didactics of Mathematics* (pp. 69-81). http://webdoc.sub.gwdg.de/ebook/e/gdm/1996/maier.pdf

Martín-Dorta, N., Saorín, J.L., & Contero, M. (2008). Development of a fast remedial course to improve the spatial abilities of engineering students. *Journal of Engineering Education*, *97*(4), 505-513. https://doi.org/10.1002/j.2168-9830.2008.tb00996.x

McGee, M.G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, *86*(5), 889–918. https://doi.org/10.1037/0033-2909.86.5.889

Mohler, J.L. (2008). Examining the spatial ability phenomenon from the student's perspective. *The Engineering Design Graphics Journal*, *72*(3), 1-15. http://gpejournal.org/index.php/EDGJ/article/view/50

National Research Council (NRC) (2006). *Learning to think spatially*. National Academies Press.

Olkun, S. (2003). Making connections: Improving spatial abilities with engineering drawing activities. *International Journal of Mathematics Teaching and Learning*, *3*(1), 1-10. http://www.ex.ac.uk/cimt/ijmtl/ijabout.htm

Olkun, S., Smith, G.G., Gerretson, H., Yuan, Y., & Joutsenlahti, J. (2009). Comparing and enhancing spatial skills of pre-service elementary school teachers in Finland, Taiwan, USA, and Turkey. *Procedia-Social and Behavioral Sciences*, *1*(1), 1545-1548. https://doi.org/10.1016/j.sbspro.2009.01.271

Ozcakir Sumen, O. (2019). A meta-synthesis about the studies on spatial skills in Turkey. *International Online Journal of Educational Sciences*, *11*(4), 23-41. https://doi.org/10.15345/iojes.2019.04.003

Patkin, D., & Dayan, E. (2013). The intelligence of observation: improving high school students' spatial ability by means of intervention unit. *International Journal of Mathematical Education in Science and Technology, 44*(2), 179-195. https://doi.org/10.1080/0020739X.2012.703335

Pellegrino, J.W., Alderton, D.L., & Shute, V.J. (1984). Understanding spatial ability. *Educational psychologist*, *19*(4), 239-253. https://doi.org/10.1080/00461528409529300

Pellegrino, J.W., & Kail, R.V. (1982). Process analyses of spatial aptitude. In R. J. Sternberg (Eds.), *Advances in the psychology of human intelligence* (Vol. I, pp. 311-365). Lawrence Erlbaum Associates.

Sarama, J., & Clements, D. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge.

Sevimli, E. (2009). *Consideration of pre-services mathematics teachers' preferences of representation in terms of definite integral within the context of certain spatial abilities and academic achievement* [Master's thesis]. University of Marmara, Turkey.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701-703. https://doi.org/10.1126/science.171.3972.701

Sisman, B., Kucuk, S., & Yaman, Y. (2021). The effects of robotics training on children's spatial ability and attitude toward STEM. *International Journal of Social Robotics, 13*(2), 379-389. https://doi.org/10.1007/s12369-020-00646-9

Sorby, S.A., & Baartmans, B.J. (2000). The development and assessment of a course for enhancing the 3-D spatial visualization skills of first year engineering students. *Journal of Engineering Education, 89*(3), 301-307. https://doi.org/10.1002/j.2168-9830.2000.tb00529.x

Tabachnick, B.G., & Fidell, L.S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Tartre, L.A. (1990). Spatial orientation skill and mathematical problem solving. *Journal for research in Mathematics Education, 21*(3), 216-229. https://doi.org/10.5951/jresematheduc.21.3.0216

Thurstone, L.L., & Thurstone, T.G. (1941). *The Chicago tests of primary mental abilities*. The American Council on Education.

Turgut, M., & Nagy-Kondor, R. (2013). Spatial visualization skills of Hungarian and Turkish prospective mathematics teachers. *International Journal for Studies in Mathematics Education, 6*(1), 168-183. https://doi.org/10.17921/2176-5634.2013v6n1p%25p

Wells, C.S., & Wollack, J.A. (2003). *An instructor's guide to understanding test reliability*. Testing and Evaluation Services, University of Wisconsin.

Wendler, C.L., & Walker, M.E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In Downing, S.M., & Haladyna, T.M. (Eds.), *Handbook of test development* (pp. 445-467). Lawrence Erlbaum Associates Publishers.

Winter, J.W., Lappan, G., Fitzgerald, W., & Shroyer, J. (1989). *Middle Grades Mathematics Project: Spatial Visualization*. Addison-Wesley.

Witkin, H.A. (1950). Individual differences in ease of perception of embedded figures. *Journal of Personality, 19*, 1-15. https://doi.org/10.1111/j.1467-6494.1950.tb01084.x

Young, C.J., Levine, S.C., & Mix, K.S. (2018). The connection between spatial and mathematical ability across development. *Frontiers in psychology, 9(1)*, 1-7. https://doi.org/10.3389/fpsyg. 2018.00755

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 98–113

https://doi.org/10.21449/ijate.1150981

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

How should citizenship education be given?: A study based on the opinions of social studies teachers

Suat Polat[1,*], Ibrahim Ozgul[2], Huseyin Bayram[3]

[1]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Social Studies Education, Agri, Türkiye
[2]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Social Studies Education, Agri, Türkiye
[3]Ağrı İbrahim Çeçen University, Faculty of Education, Department of Social Studies Education, Agri, Türkiye

ARTICLE HISTORY

Received: July 29, 2022

Accepted: Mar. 02, 2023

Keywords:
Citizenship education,
Social studies education,
Social studies teachers,
Teachers' opinions.

Abstract: In this study, it was aimed to perform an in-depth examination of the opinions of social studies teachers on how citizenship education should be given. Phenomenology design, one of the qualitative research methods, was used in the research. The study group of the research was formed with 17 social studies teachers working in different regions of Turkey, having different professional seniority and also different genders, by using maximum diversity sampling. The data of the study were collected through a semi-structured interview form developed by the researchers. Descriptive analysis was used to analyze the data. As a result of the research, it was concluded that the perception of social studies teachers regarding the phenomenon of citizenship varied. It was determined that social studies teachers considered the main purpose of social studies as citizenship education and also as a course aimed at raising individuals that the society needs. In the study, it was also concluded that the teachers were of the opinion that appropriate content should be used, different teaching methods-techniques should be employed, and value education should be emphasized while giving citizenship education in the social studies course. Various suggestions were made based on the results of the research.

1. INTRODUCTION

Education is the process of bringing about a deliberate and desired change in an individual's behavior through his/her own experiences (Ertürk, 1973). The sum of the processes in which an individual acquires skills, orientation and behavioral patterns in the society he/she lives in is called education (Demirel & Kaya, 2002). Human is at the center of education. In other words, both the subject and the object of education are humans. All activities carried out through education are carried out with the aim of changing human behavior towards the desired direction (Sönmez, 1994).

Human is a social being and has to live in a society. Living in peace and security in a society is possible with the establishment of social order. Accordingly, some written and unwritten rules have been set to establish social order in the historical process, and states have been established

*CONTACT: Suat Polat ✉ spolat@agri.edu.tr Agri Ibrahim Cecen University, Faculty of Education, Department of Social Studies Education, Agri, Türkiye.

e-ISSN: 2148-7456 /© **IJATE** 2022

(Bilge, 1990). While some rights and responsibilities are given to people in order to maintain the social order, some of their freedoms are also restricted to maintain this order. Unlimited freedoms do not seem possible in the social structure as they affect the freedoms of other people. Accordingly, it is important for all people to comply with social rules, whether written or unwritten (Ereş, 2015).

In the historical process, societies have aimed to raise good and active citizens who have internalized the political systems and social structures of their countries. The reason for this is based on the view that it is possible to ensure the continuity of countries and to unite differences in common purposes by instilling citizenship awareness in people. Accordingly, in recent years, education systems have made an intense effort to serve the purpose of raising active and effective citizens. It is a result of these efforts that the socialist understanding built on the view that the individual is responsible to society forms the basis of citizenship education. This approach tries to establish a balance between social rights and responsibilities and individualism. At this point, citizenship education both tries to strengthen the social, political and moral duties and responsibilities of young people and functions as a social control mechanism on individuals. Citizenship knowledge, skills and values of individuals are tried to be strengthened through citizenship education (Wood, 2010). The definition of citizenship education is generally made on the basis of the following content:

To bring young people in the roles and responsibilities they should have in order to grow up as effective citizens. Education experts define this process as "citizenship education". In citizenship education, subjects such as history, geography, economy, law, politics, linguistics, environmental knowledge and international studies, social studies, life skills and moral education have an important place. The content related to these fields and subjects is taught to primary and secondary school students through social studies education. Citizenship education is currently very up-to-date in many countries. Educational approaches that emerged in the new century are meticulously focused on how to perform citizenship education. As a matter of fact, it is important to prepare young people quickly for the difficulties, uncertainties and changing world conditions (Ichilov, 1998). The majority of IRCAF countries (Australia, Canada, England, France, Germany, Hungary, Italy, Japan, South Korea, Netherlands, New Zealand, Singapore, Spain, Sweden, Switzerland, USA, Hong Kong, Ireland) that give importance to citizenship education make important reforms in school curricula in which they develop projects for quality citizenship education. At the end of the 20th century, the level and speed of global change brought thematic studies to the fore. For example, the content and implementation approach of citizenship education in England is being reshaped in accordance with the requirements of the time. In this context, thematic study areas are formed for citizenship education (Crick, 1998; Kerr, 1999). Development of citizenship education in Turkey it happened in a different way (Üstel, 2005). After the World War I, the Ottoman Empire collapsed. Republic of Turkey established on October 29, 1923. After the establishment of the republic reforms have been carried out in many areas. One of these ares is education (Ata, 2006). In the 1926 and 1936 curricula, it was aimed to strengthen national feelings and to consolidate the reforms made (Kuş, 2014). In the 1962 draft curriculum instead of the history, geography and civics course taught in the fourth and fifth grades, a new course called society and country studies took place. The name of social studies course in Turkey was included in the 1968 curriculum for the first time. The 1968 program was repealed in 1985. In 1985, social studies education was ended in secondary schools. Instead of social studies national geography and national history courses were took place. In 1992, the citizenship course has been placed. After the eight-year compulsory and uninterrupted primary education came into effect in 1997, national history and national geography courses have been removed from primary schools. Citizenship education started to be given through the social studies course (Şen, 2019). The social studies course, which is taught at primary and secondary school levels in our country, has started to be taught within the framework of thematic study areas in accordance with the constructivist approach since 2005. Citizenship education is one of these themes. This course helps children grow up as active citizens who successfully integrate into the society and perform the socialization process.

Socialization of citizens is considered important in a democratic society. This socialization process should start in childhood and primary schools. In this period, social studies course has an important mission both in terms of character, moral and values education of children and in gaining citizenship knowledge, skills and values (Althof & Berkowitz, 2006). While citizenship is performed by binding people to the state with a legal bond, raising an active citizen who is aware of his/her rights and responsibilities is possible with an effective citizenship education. Citizenship education serves to establish a healthy social structure by strengthening the bonds between the individual and the society. In this direction, the social studies course, which serves to ensure the adaptation of individuals to social life by internalizing democratic principles and values, helps individuals to apply the citizenship knowledge, skills and values they have acquired within the scope of citizenship education in their daily life (Merey, 2009).

The special objectives of social studies curriculum such as raising the individuals as citizens of the Republic of Turkey who love their homeland and nation, know and use their rights, fulfill their responsibilities, and have national consciousness, and having them understand the historical processes of the concepts of human rights, national sovereignty, democracy, secularism, republic and their effects on today's Turkey and organize their lives according to democratic rules (MoNE, 2018) clearly show the importance of citizenship education in this course. When the literature is reviewed, it is possible to find studies on citizenship education (Şen, 2019; Bıçak & Ereş, 2018; Önal et al., 2018; Önal et al., 2017; Şentürk et al., 2017; Gürel, 2016; İpek & Karataş, 2015; Deveci & Selanik Ay, 2014; Memişoglu, 2014; Ichilov, 2013; Geboers et al., 2013; Merey et al., 2012; Hablemitoğlu & Özmete, 2012; Kennedy, 2012; Hebert & Sears, 2001; Kerr, 1999). Considering both the relevant literature and social studies curriculum, it is observed that citizenship education is very important for this course. The fact that the social studies course was started to be taught with the aim of raising active citizens further reinforces this importance. For all these reasons, the opinions of the social studies teachers, who teach this course, on how citizenship education should be gain importance. It is thought that this study will contribute to the existing literature by revealing the views of social studies teachers on how citizenship education should be.

In this study, it was aimed to perform an in-depth examination of the opinions of social studies teachers on how citizenship education should be given. For this purpose, answers were sought to the following questions:

- What is the perception of social studies teachers about the phenomenon of citizenship?
- What are social studies teachers’ views about the role of social studies course in citizenship education?
- What are social studies teachers’ views on the way to be followed while giving citizenship education in social studies course?

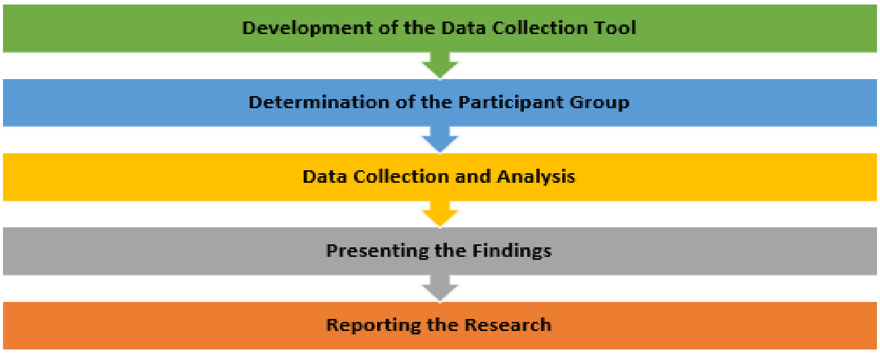
2. METHOD

Qualitative research methods were used in the study. Qualitative research is a method that examines events, facts and situations in their own conditions and forms and aims to collect in-depth information (Glesne, 2016). The reason for using the qualitative research method in the study is to collect information by examining the views of the participants on the research topic in depth. This research was conducted within the scope of the permission obtained by the decision of scientific research ethics committee of Ağrı İbrahim Çeçen University, dated 23.02.2022 and numbered 56.

2.1. Design

In this study, in which the views of social studies teachers on how citizenship education should be given were researched, phenomenology, one of the qualitative research designs, was used. Phenomenology focuses on people's perceptions and views that develop depending on their lives (Creswell & Poth, 2018). The reason for using the phenomenology design in this study is to reveal the experiences of the social studies teachers participating in the research on citizenship education and their views on how citizenship education should be given in the social studies course by examining in detail. The stages followed in the research conducted using the phenomenology design are shown in Figure 1.

Figure 1. *Stages followed in the research process.*


Development of the Data Collection Tool
Determination of the Participant Group
Data Collection and Analysis
Presenting the Findings
Reporting the Research

As seen in Figure 1, a gradual process was followed in the research. Five stages were followed, from the determination of the data collection tool to the reporting of the process, in the research. In the research process, the research fit matrix (Kaya & Bayram, 2021) was used to control the compatibility between the variables of the research.

2.2. Study Group

Maximum diversity, one of the purposive sampling methods, was used in the study. The main purpose of maximum diversity sampling is to reflect the diversity of the individuals participating in the study at the maximum level. The aim of maximum diversity sampling is not to generalize, but to determine whether there are shared phenomena among the diverse situations (Yıldırım & Şimşek, 2016). The reason for using maximum diversity sampling in this study was that the social studies teachers who worked in different geographical regions and had different professional seniority and different genders were planned to be included in the study. Information about the social studies teachers participating in the research is presented in Table 1.

In Table 1, it is seen that 47.05% of the participants in the participant group were female and 52.95% were male. In the table, it is observed that 23.53% of the participants were 25-35 years old, 41.17% were 36-45 years old, 17.65% were 46-55 years old, and 17.65% were over 55. In addition, the table shows that 29.41% of the participants worked in the Marmara region, 23.3% in Central Anatolia, 17.65% in the Mediterranean, 11.76% in the Black Sea, and 17.65% in the Southeastern Anatolia region. Since social studies teachers in Eastern Anatolia and Aegean regions could not be reached, participants from these two regions were not included in the study. Within the scope of the information in the table, it is understood that 23.53% of the participants had a professional seniority of 1-5 years, 41.17% of them 6-10 years, 17.65% of them 11-15 years and 17.65% of them had a professional seniority of more than 15 years.

Table 1. *Information on social studies teachers included in the study.*

Variable	Feature	f	%
Gender	Female	8	47.05
	Male	9	52.95
	Total	17	100
Age	25-35	4	23.53
	36-45	7	41.17
	46-55	3	17.65
	55+	3	17.65
	Total	17	100
The Region of Duty	Marmara Region	5	29.41
	Central Anatolia Region	4	23.53
	Mediterranean Region	3	17.65
	Black Sea Region	2	11.76
	Southeastern Anatolia Region	3	17.65
	Total	17	100
Professional Seniority	Between 1-5 years	4	23.53
	Between 6-10 years	7	41.17
	Between 11-15 years	3	17.65
	Over 15 years	3	17.65
	Total	17	100

The names of the participants in the research were kept confidential within the scope of ethical rules, and a code name was formed for each participant. The code names in question were formed by adding a number to the letter "T", which is the first letter of the word teacher, according to the order of the interview. In this context, the first interviewed participant was included in the research with the code name T1, the second participant T2, and the third participant T3.

2.3. Data Collection Tools and Data Collection

The data of the study were collected through a semi-structured interview form developed by the researchers. During the development of the interview form used in collecting the data of the research, the views of the experts on qualitative research were obtained three times. The interview form developed by the researchers was first sent to two experts, and they were asked to express their opinions in terms of its suitability to the subject of the research. Within the scope of the feedback received, all the questions in the interview form were rearranged. Then it was presented to the experts again for their opinions. At this stage, feedback was received, indicating that some of the questions in the form were repeated. Depending on the feedback received, some of the questions in the form were combined, and the number of questions was decreased. Finally, the opinion of a different expert was asked and the final version of the interview form was formed. Based on the opinions of the experts, it was decided that the interview form was appropriate for this research. Therefore, it was not necessary to conduct a pilot application for the interview form.

The data of the research were collected face-to-face and web-based. The fact that the participants in the study were working in different regions of Turkey was determined as the limitation of the study, and the participants were interviewed using the teleconference method, where the researchers could not find the opportunity to collect data face-to-face. In this context, while face-to-face interviews were held with the participants in the South-East Anatolia region, teleconferences were held with the participants in the other regions. Voice recordings were made with the permission of the participants, both in face-to-face interviews and in interviews made by teleconference method. These recordings were transcribed in electronic environment and arranged for the analysis phase.

2.4. Analysis of Data

The data obtained through the interview forms were analyzed through descriptive analysis, one of the qualitative data analysis methods. The descriptive analysis includes performing the analysis process within the scope of predetermined themes (Yıldırım & Şimşek, 2016). The reason why descriptive analysis was used in the analysis of the data of this research is that the coding was done according to the themes were predetermined.

The data transcribed in the computer environment were analyzed in four stages. In the first stage, the data were read superficially to have an idea about the data set. In the second stage, the data were read again and the first coding was done. In the third stage, the main coding was done. In the fourth stage, all the codes determined were brought together and themes were formed. In order to ensure the reliability of the research during the analysis process, the fourth part of the data set and the codes and themes extracted from the said part were sent to two different experts and they were asked to code. By comparing the codes of the experts with the codes of the researchers, the discrepancies were resolved and the themes were finalized. [Reliability=Agreement / (Agreement+Disagreement)] formula developed by Miles and Huberman (1994) was used to determine the reliability of the coding. As a result of the calculation, the reliability value was found to be 92%. The fact that the result of the abovementioned formula is over 70% indicates that the analysis is reliable (Miles & Huberman, 1994).

2.5. Research Ethics and Validity and Reliability

In the research, the following procedures were carried out within the scope of scientific research ethics:

- Before starting the research, ethics committee permission was obtained from a university's scientific research ethics committee to conduct the research.
- The scope of the research was clearly explained to the participants included in the research, and it was ensured that the participants were informed about the research.
- The recordings taken while collecting the research data were used only for this research.
- The names of the participants in the research, the cities they resided in and the institutions they worked in were kept confidential.

In order to ensure the validity and reliability of the research, the following procedures were carried out:

- In the process of designing the semi-structured interview form used in the collection of research data, the opinions of experts who are competent in qualitative research were taken.
- The data collected during the research process were filed based on the original forms.
- The opinions of experts were taken regarding one-fourth of the codes and themes formed during the qualitative analysis process.

The reliability formula developed by Miles and Huberman (1994) was used to determine the reliability of the coding made during the analysis process.

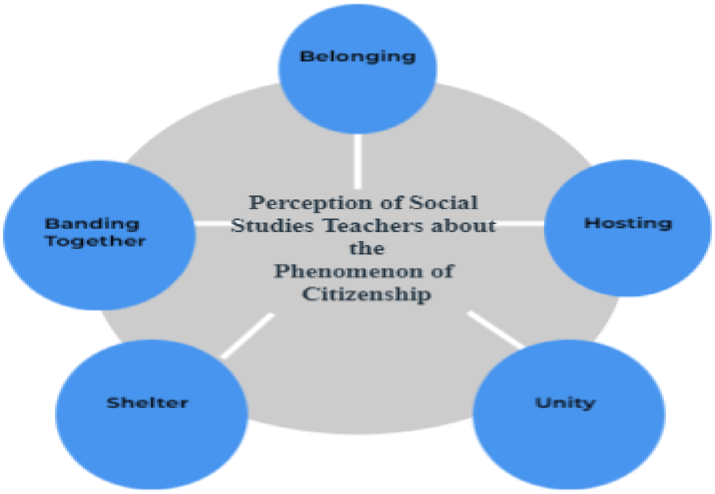
3. RESULT

In the research, the views of social studies teachers on how citizenship education should be given were examined in depth. In this context, the perception of the phenomenon of citizenship of the social studies teachers, their views on the role of social studies course in citizenship education and their views on what path to follow in citizenship education in social studies course were examined.

3.1. Findings Obtained in the Scope of the Perception of Citizenship of the Social Studies Teachers

In the study, the perception of social studies teachers about the phenomenon of citizenship was researched. The findings obtained in this context are shown in Figure 2.

Figure 2. *Findings reached within the scope of the perception of citizenship of the social studies teachers.*


Belonging
Perception of Social Studies Teachers about the Phenomenon of Citizenship
Banding Together
Hosting
Shelter
Unity

As seen in Figure 2, the findings reached in the research were combined under the themes of belonging, having a home, unity, shelter and banding together. The findings within the scope of these themes are as follows:

It was determined that some of the participants in the research perceived the phenomenon of citizenship as an element of belonging. One of the most striking expressions providing the aforementioned finding was presented by T3. T3 expressed his/her perception on the subject by saying *"Citizenship means that a person belongs to a place, a country, a land."* T10 also expressed his/her opinion with similar sentences. T10 said *"Every person wants to belong somewhere. A human being wants to belong to a community. Here, I think citizenship means belonging to a community."* and revealed that he/she perceived the phenomenon of citizenship as being a member of a community. Another participant who associated citizenship with belonging was T6. This participant (T6) said *"Citizenship is derived from the word city (state in ancient times). So, it means belonging to a country. It means internalizing a homeland. In my opinion, citizenship means being a part of a homeland."* and embodied his perception on the subject. T6 continued *"In addition, citizenship also enables people to have an identity by belonging to a place. That is why it is what makes us human.''* and expressed the importance he/she attributed to the phenomenon of citizenship.

It was determined that one of the teachers participating in the research perceived citizenship as having a home. The teacher mentioned is T5. T5 presented his/her perception in this context with the following words:

> *"Homeland means the home of a person. It means the house of his/her ancestor. Citizenship then means being the owner of that home. The holiest place of a person is his/her home. He/she is born at home, he/she feeds himself/herself at home. He/she becomes happy at home, sad at home. He/she dies at home. I mean, the homeland is the place where people do everything. For this reason, I will say that homeland is the home of a person. Home is also very important. In fact, one's home is one of the most important things in life."*

It was determined that the perception of citizenship of some participants and the phenomenon of citizenship were based on unity. As a matter of fact, T8 said *"Citizenship means unity. If people are united, they become citizens. They become brothers. So they develop their homeland. That is why being united is the root of being a citizen, in my opinion."* and embodied the phenomenon of citizenship. T9 also presented his/her perception of the phenomenon of citizenship in a similar way. T9 said *"You know, when we were kids, we used to sing in games - one for all, all for one - I think citizenship is something like that. Unity means power. Citizenship is also the source of power. Therefore, citizenship is equivalent to being united."* and s/he expressed that s/he perceived citizenship as unity.

It was found that the participant with the code T4 in the research perceived citizenship as a shelter. The striking statements of T4 providing the aforementioned finding are as follows:

> *"People always look for a shelter in life. Citizenship is that shelter. No matter what a person experiences, knowing that there is a shelter in which he/she can be safe makes one feel strong. When a person feels weak or feels lonely, he/she can enter this shelter and rest. I think the most beautiful shelter is citizenship. A shelter where you can enter unquestioningly."*

It was observed that some participants in the study expressed the phenomenon of citizenship in the form of banding together. For example, one of the participants, T1, embodied his/her thoughts as follows: *"Doesn't citizenship mean being a stakeholder of the homeland? Yes. So, citizenship means that all stakeholders band together."* Another participant, T14 said *"In my opinion, citizenship is banding together.''* and T17 said *"Citizenship means banding together. What else could it be? To unite. Being a part of a whole."*

Within the scope of the findings in the research, it was determined that the perception of social studies teachers about the phenomenon of citizenship varied. On the other hand, it was determined that all social studies teachers participating in the research attributed deep and comprehensive meanings to the phenomenon of citizenship and also considered it important for human life.

3.2. Findings Obtained Within the Scope of Social Studies Teachers' Opinions on the Role of Social Studies Course in Citizenship Education

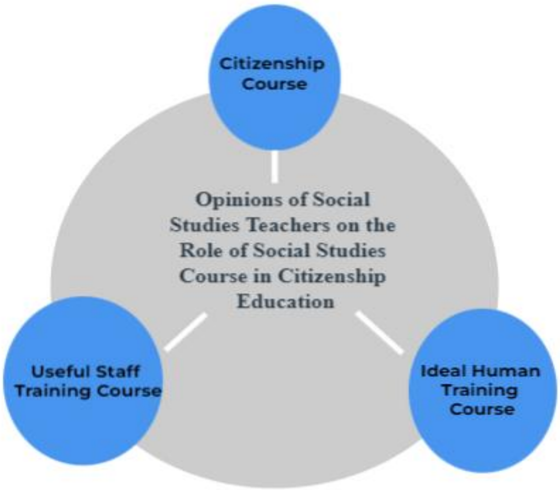
In the research, the opinions of social studies teachers on the role of social studies course in citizenship education were investigated. The findings are shown in Figure 3. As seen in Figure 3, the findings reached within the scope of the opinions of social studies teachers on the role of social studies course in citizenship education were brought together under the themes of citizenship course, ideal human training course, and useful staff training course. The findings within the scope of these themes are as follows:

It was determined that most of the participants in the study expressed the role of social studies course in citizenship education as a course for citizenship education in primary and secondary school programs. One of the most striking opinions in this context was expressed by participant T13 as follows:

> *"Social studies is already the citizenship course itself. Therefore, there is no need to even think about the role of this course (social studies) in citizenship education. When we look back at history, the reason for the emergence of social studies is citizenship education. It is still so today. It will also be*

> *the same in the future. When it comes to citizenship course, everyone immediately thinks of social studies. This has always been the case both in Turkey and in the world. It will always be in this way."*

Figure 3. *Findings reached within the scope of social studies teachers' opinions on the role of social studies course in citizenship education.*


Citizenship Course
Opinions of Social Studies Teachers on the Role of Social Studies Course in Citizenship Education
Useful Staff Training Course
Ideal Human Training Course

The participant with the code T2 also expressed a similar opinion. T2 said *"Don't we already teach social studies to raise citizens? Yes, we do. Then, the importance and role of social studies should be considered in terms of citizenship course."* T2 also said *"Open and look at social studies books. Everything is about citizenship education. No matter what you look at, you always encounter citizenship issues in the books."* and he evaluated the social studies textbooks in terms of citizenship education. Another participant, T7 said *"I think social studies is already citizenship education. It is citizenship education itself. In fact, its origin is in citizenship education."*

It was determined that some participants expressed the role of social studies course in citizenship education as social studies being the ideal human training course. In fact, T11 said *"I think it is enough to look at the definition of social studies to answer this question. What does it say in the definition? It says that social studies is an ideal citizen training course. I think this is the best answer.''* T11 also said *"What do we teachers (social studies teachers) do? We teach to raise quality, rightminded, good citizens. This is the role of social studies."* and embodied his/her opinion on the subject within the scope of the mission undertaken by social studies teachers. The participant with the code T15 expressed his/her opinion on the role of social studies course in citizenship education in a similar way. T15 said *"This (the role of social studies course in citizenship education) is very clear to me. Social studies is the lesson of raising the people desired by the society and the state. This is very clear."* and he/she expressed his/her point of view on the subject very clearly.

T12, one of the participants in the research, expressed the role of social studies course in citizenship education by expressing social studies in the form of a useful staff training course. The statements of T12 in this context are as follows:

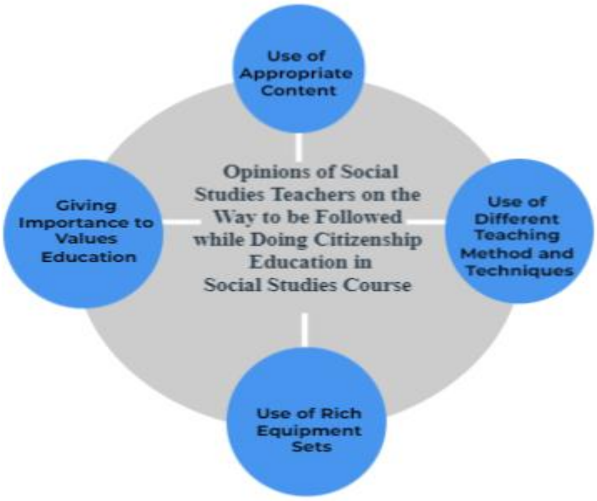
> *"In my opinion, social studies is a staff training course. Why do I say it is a staff training course? Because we always call it an active citizen training course for social studies. Who is this active citizen? I think I need to clarify. For example, we provide students with a lot of skills. We don't just bring them in knowledge or value. We provide a lot of skills that can be used in life. That is why I say that. Social studies is a staff training course. In the future, these students will deal with different fields. They will be the staff of different fields. Here we train staff for different fields. I mean, we are preparing the infrastructure."*

According to the findings obtained in the research, it was determined that social studies teachers considered social studies course as the basic course of citizenship education. In addition, it was revealed that social studies teachers thought of social studies course as a tool to form the individual profile desired by the society and the state, and also as a course that equips primary and secondary school students with various skills for their future orientation.

3.3. Findings Obtained Within the Scope of Social Studies Teachers' Opinions on the Way to Be Followed While Conducting Citizenship Education in Social Studies Course

In the research, the opinions of social studies teachers on the way to be followed while giving citizenship education in social studies course were examined. The findings are shown in Figure 4.

Figure 4. *Findings obtained within the scope of the opinions of social studies teachers on the way to be followed in while giving citizenship education in social studies course.*


Use of Appropriate Content
Giving Importance to Values Education
Opinions of Social Studies Teachers on the Way to be Followed while Doing Citizenship Education in Social Studies Course
Use of Different Teaching Method and Techniques
Use of Rich Equipment Sets

As can be seen in Figure 4, the findings obtained within the scope of the opinions of social studies teachers on the way to be followed for citizenship education in the social studies course were combined under the themes of the use of appropriate content, the use of different teaching methods-techniques, the use of rich equipment sets, and the emphasis on value education. The findings within the scope of these themes are as follows:

It was determined that some of the teachers who participated in the research emphasized that appropriate course contents should be used while giving citizenship education in the social studies course. For example, T16 expressed his/her opinion on this issue with the following words:

> *"Citizenship education is extremely important. Therefore, it is necessary to use the right content. We normally only use the books of the Ministry of National Education. I think this is not enough. I find the content of the textbooks insufficient in this respect (citizenship education). I think that the content on citizenship should be increased. Especially in the fourth grade. Because the student in the fourth grade has a fresher brain. He/She gets what you give. Therefore, the content should be enriched in the fourth-grade books."*

The participant with the code T3 also used similar expressions. T3 said *"The question of what we will teach in citizenship education is important. What to teach depends on the content. This means that while giving citizenship education, it is necessary to prepare and use content suitable for raising citizens.''* and he/she drew attention to the importance of the content of citizenship education in the social studies course.

It was determined that some of the participants had the opinion that different teaching methods-techniques should be used while giving citizenship education in the social studies course. The most striking opinion in this context was expressed by T6. T6 expressed his/her point of view on the subject as follows:

> *"It is necessary to use ways that work in education. A good teacher should know well which method to use. First, he/she needs to know the student. What are the student's needs? How is the student's character? How can a student be taught? We need to know all of these. For example, there are some students. They memorize. No matter what you do, you cannot make them stop memorizing. For example, another type of student learns as s/he speaks in class. You need to know all of these. Why? Because if you are going to give citizenship education, you need to know these."*

T9 expressed his/her opinion in this context as follows: *"Diversity is good when teaching. Different techniques can be used. It should not always be direct instruction. Especially not in citizenship education.''* T9 also said *"For example, we can use techniques in which the student is more involved. Thus, we save the student from memorization.''* and he/she emphasized that active learning models can be used while giving citizenship education in social studies course.

It was determined that a few teachers in the research had the opinion that rich equipment sets should be used while giving citizenship education in social studies course. T14, one of the aforementioned participants, said, *"If quality education is to be given, all kinds of materials must be available. If there is a lack of material while explaining a subject, effective teaching cannot be achieved. The situation is the same in citizenship education. If a subject is to be taught, we should have enough material.''* and he/she expressed his/her opinion that experiencing an effective education process is directly proportional to having sufficient tools and equipment. As another participant, T8 said *"You must have tools that you can use in the classroom so that the education can achieve its purpose. I am not talking about a smart board or a projector. With these, very limited activities can be done."* and he/she expressed her opinion on the subject. He/she also said *"If you are going to talk about an event, you must have miniatures, topographical maps, a hall and costumes for re-enactment."* and embodied the tools that he/she considered to be used during citizenship education in the social studies course.

It was determined that two participants had the opinion that values education should be emphasized while giving citizenship education in the social studies course. One of the aforementioned participants, T1, expressed his/her point of view on the subject as follows:

> *"Citizenship education means value education. We give a lot of value education in social studies. As a matter of fact, good value education is given in social studies. However, this is not enough. I think that a good citizen should have all material and moral values. In my opinion, the main way to follow in citizenship education is to focus on values education. The more the students internalize our values, the more effective citizens they can become."*

The other participant who thought that value-oriented citizenship education should be taken as a basis in the social studies course was T10. T10 said *"Social studies is a citizenship education course anyway. In the social studies course, we teach students everything that a citizen should have. But for good education, we should emphasize our values.''* T10 also said, *"There is a reason why I say that we should emphasize values. A citizen must have values above all else. In my opinion, this is the main feature that a person should have."* and he/she based her opinion on striking statements.

Within the scope of the findings, it was revealed that the social studies teachers included in the research were of the opinion that the content suitable for citizenship education should be used while conducting citizenship education in the social studies course. In addition, it was determined that the participants thought that the teaching methods-techniques should be diversified in the processes of citizenship education and at the same time, they should have the necessary tools and equipment. On the other hand, it was determined that social studies teachers were of the opinion that effective citizenship education can be possible with comprehensive values education processes.

4. DISCUSSION and CONCLUSION

Citizenship education is given through social studies course at primary and secondary schools, as it is related to social life. In fact, knowledge, skills and values related to social life are the subject of social studies course. Considering the fact that the knowledge, skills and values that the ideal citizen should have gained through the social studies course, the issue of how citizenship education should be given in the social studies course gains importance. The results obtained in this research, which was based on the opinions of social studies teachers, were discussed within the scope of the results of similar studies in the relevant literature and presented below.

In the study, it was concluded that social studies teachers had a deep perception of citizenship. As a matter of fact, it was determined that the teachers attributed citizenship phenomenon meanings such as belonging, unity, banding together, and that they also considered citizenship as one of the most important needs for human beings. In the literature review, studies with results similar to those of this research were found. For example, the participants included in the study by Kadıoğlu et al. (2016) stated the phenomenon of citizenship as language unity, religious unity, common culture and laws that bind everyone. The participants of the study conducted by Martin and Chiodo (2007) associated the phenomenon of citizenship with expressions of benevolence, adherence to social rules, and being respectful. As a result of their study with prospective teachers, Değirmenci and Eskici (2019) determined that the participants associated citizenship with being responsible in general. Malkoç and Ata (2021), on the other hand, identified that social studies teachers considered citizenship as a way of establishing organic bonds with history and culture.

In the present study, it was concluded that social studies teachers considered social studies as a course that is directly aimed at citizenship education and at the same time helping to raise qualified individuals. In the literature review carried out to discuss the abovementioned result, studies with similar results were found. Kuş and Aksu (2017), in their study with social studies teachers, determined that teachers considered social studies as a course whose main purpose is citizenship education. Memişoğlu (2014) concluded that social studies teachers considered social studies as a citizen education course and also as a course that adds knowledge, skills and values. In their study with prospective teachers, Deveci and Selanik Ay (2014) determined that the participants considered social studies as a course that facilitates life in citizen education and provides social order. Karasu Avcı et al. (2020) found that social studies teachers thought that social studies course adds value to students and raises effective citizens.

In the present study, it was concluded that the social studies teachers were of the opinion that the content suitable for citizenship education should be used while providing citizenship education in the social studies course. In the study, it was also determined that social studies teachers thought that different teaching methods-techniques should be used in citizenship education and that they should have the necessary tools and materials. In addition, in the study, it was determined that social studies teachers were of the opinion that values education should be emphasized for effective citizenship education. Similarly, Bıçak and Ereş (2018) concluded in their study that teachers emphasized the importance of content in citizenship education. Memişoğlu (2014) determined that social studies teachers tried to provide diversity in classroom and out-of-class activities during the citizenship education process. Wilkins (2003) revealed in his study with teachers that contrary to this research, teachers limited citizenship education to classroom activities only. Pederson and Cogan (2000) concluded that citizenship education was carried out only with the direct instruction technique. Kuş and Aksu (2017) revealed that social studies teachers considered social studies as a course that helps to raise effective citizens by adding value.

As a result of the present research, it was concluded that the perception of social studies teachers regarding the phenomenon of citizenship varied. On the other hand, it was determined that the social studies teachers considered the main purpose of social studies as citizenship education and at the same time, they considered it as a course aimed at raising individuals that society needs. In the study, it was also concluded that the teachers were of the opinion that appropriate content and different teaching methods-techniques should be used, and value education should be emphasized while giving citizenship education in the social studies course.

Various recommendations were made depending on the results of the research. These recommendations are listed below:

- In-service training on how to give citizenship education can be provided for social studies teachers.
- Social studies teachers can be provided with opportunities to use different teaching methods and techniques in citizenship education.
- Social studies textbooks and the content of the social studies course curriculum can be revised within the scope of compliance with citizenship education. The abovementioned books and the curriculum can be enriched in terms of citizenship education based on values education.
- Further researches can investigate the opinions of social studies teachers on citizenship education with bigger participant groups.
- Further researches can be made on the opinions of social studies teachers on citizenship education with quantitative methodology.
- Further researches can be made on the opinions of social studies teachers on citizenship education with mixed methods.

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. This study was conducted within the scope of the permission obtained by the decision of scientific research ethics committee of Ağrı İbrahim Çeçen University, dated 23.02.2022 and numbered 56.

Authorship Contribution Statement

The contribution rate of all authors to the article is equal.

Orcid

Suat Polat https://orcid.org/0000-0001-9286-8840
Ibrahim Ozgul https://orcid.org/0000-0002-5325-8326
Huseyin Bayram https://orcid.org/0000-0001-6065-8865

REFERENCES

Althof, W., & Berkowitz, M.W. (2006). Moral education and character education: Their relationship and roles in citizenship education. *Journal of Moral Education, 35*(4), 495-518. https://doi.org/10.1080/03057240601012204

Ata, B. (2006). Sosyal bilgiler öğretim programı [Social studies curriculum]. In C. Öztürk (Ed.), *Hayat bilgisi ve sosyal bilgiler öğretimi: Yapılandırmacı bir yaklaşım [Life studies and social studies teaching: A constructivist approach]* (p.71-83). Pegem Academy Publication.

Bıçak, İ., & Ereş, F. (2018). Öğretmenlerin vatandaşlık eğitimi ve vatandaşlık eğitimi sürecine yönelik görüşleri [Teachers' views on citizenship education and citizenship education process]. *The Journal of Turkish Educational Sciences, 16*(2), 257-279. Retrieved from https://dergipark.org.tr/en/pub/tebd/issue/41575/457068

Bilge, N. (1990). *Hukuk başlangıcı, hukukun temel kavram ve kurumları [Beginning of law, basic concepts and institutions of law]*. Turhan Publications.

Creswell, J.W., & Poth, C.N. (2018) *Qualitative inquiry and research design choosing among five approaches*. Sage.

Değirmenci, Y., & Eskici, B. (2019). Öğretmen adaylarının etkin vatandaşlık algılarının incelenmesi [Examination of prospective teachers' perceptions of active citizenship]. *OPUS International Journal of Society Researches, 11*(18), 232-256. https://doi.org/10.26466/opus.537074

Demirel, Ö., & Kaya, Z. (2002). *Öğretmenlik mesleğine giriş [Introduction to the teaching profession]*. Pegem Academy Publication.

Deveci, H., & Selanik Ay, T. (2014). Vatandaşlık eğitimi bakımından sosyal bilgilerin toplumsal gücü [The social power of social studies in terms of citizenship education]. *Anadolu University Journal of Social Sciences, (Special Issue)*, 97-109. Retrieved from https://app.trdizin.gov.tr/publication/paper/detail/TWpZeE1qSXlNZz09

Ereş, F. (2015). Vatandaşlık eğitimi ve karakter eğitimi politikalarının değerlendirilmesine yönelik nitel bir çalışma [A qualitative study on the evaluation of citizenship education and character education policies]. *Mehmet Akif Ersoy University Journal of Education Faculty, 1*(36), 120-136. Retrieved from https://dergipark.org.tr/en/pub/maeuefd/issue/19409/206393

Ertürk, S. (1973). *Eğitimde program geliştirme [Curriculum development in education]*. Yelkentepe Publication.

Geboers, E., Geijsel, F., Admiraal, W., & Ten Dam, G. (2013). Review of the effects of citizenship education. *Educational Research Review, 9*, 158-173. https://doi.org/10.1016/j.edurev.2012.02.001

Glesne, C. (2016). *Becoming qualitative researchers: An introduction*. Pearson Publishing.

Gürel, D. (2016). Sınıf ve sosyal bilgiler öğretmenlerinin ilkokul 4. sınıf insan hakları, yurttaşlık ve demokrasi dersine yönelik görüşlerinin karşılıklı olarak incelenmesi [A reciprocal review on opinions of form teachers and social studies teachers on the human rights, citizenship and democracy course provided at the 4th grade of primary school]. *Ahi Evran University Journal of Kırşehir Education Faculty, 17*(3), 641-660. Retrieved from https://dergipark.org.tr/en/pub/kefad/issue/59425/853520

Hablemitoğlu, Ş., & Özmete, E. (2012). Etkili vatandaşlık eğitimi için bir öneri [A suggestion for effective citizenship education]. *Journal of Ankara Health Sciences, 1*(3), 39-54. Retrieved from https://dergipark.org.tr/en/pub/ausbid/article/445606

Hébert, Y.M., & Sears, A. (2001). Citizenship education. Canadian Education Association.

Ichilov, O. (2013). *Citizenship and citizenship education in a changing world*. Routledge.

Ichilov, O. (ed.) (1998). *Citizenship and citizenship education in a changing world*. Woburn Press.

İpek, S.O.M., & Karataş, H. (2015). Türkiye'de vatandaşlık eğitimi üzerine bir inceleme [An investigation on citizenship education in Turkey]. *Uşak University Journal of Educational Research, 1*(1), 33-50. https://doi.org/10.29065/usakead.232402

Kadıoğlu, A., Keyman, F., & Çakmaklı, D. (2016). *Vatandaşlık araştırması bulgular raporu [Citizenship research findings report)*. (Konda/IPM). Istanbul Policy Center, Sabancı University. Retrieved from https://konda.com.tr/wpcontent/uploads/2017/02/2016_03_VatandaslikArastir masi.pdf

Karasu Avcı, E., Faiz, M., & Turan, S. (2020). Etkili vatandaşlık eğitiminde değerler eğitimi: Sosyal bilgiler öğretmenlerinin düşünceleri [Values education in effective citizenship education: Thoughts of social studies teachers]. *Journal of Values Education, 18*(39), 263-296. https://doi.org/10.34234/ded.655916

Kaya, E., & Bayram, H. (2021). Utilization of the research compliance matrix in educational research design and evaluation: A design based research. *International Journal of Education Technology and Scientific Researches, 6*(15), 887-944. http://dx.doi.org/10.35826/ijetsar.325

Kennedy, K. (2012). *Citizenship education and the modern state.* Routledge.

Kerr, D. (1999). *Re-examining citizenship education: The Case of England.* NFER.

Kuş, Z. (2014). What kind of citizen? Analysis of social studies curriculum in Turkey. *Citizenship, Social and Economics Education, 13*(2), 132-145. https://doi.org/10.2304/csee.2014.13.2.132

Kuş, Z., & Aksu, A. (2017). Vatandaşlık ve vatandaşlık eğitimi hakkında sosyal bilgiler öğretmenlerinin inançları [Social studies teachers' beliefs on citizenship and citizenship education]. *International Journal of Turkish Education Sciences, 2017*(8), 18-41. Retrieved from https://dergipark.org.tr/en/pub/goputeb/issue/34591/382003

Malkoç, S., & Ata, B. (2021). Sosyal bilgiler öğretmenlerinin vatandaşlık tiplerinin belirlenmesi [Determination of the citizenship types of social studies teachers]. *Ankara University Journal of Faculty of Educational Sciences (JFES), 54*(2), 459-496. https://doi.org/10.30964/auebfd.897458

Martin, L.A, & Chiodo, J.J. (2007). Good citizenship: What students in rural schools have to say about it? *Theory and Research in Social Education, 35*(1), 112-134. https://doi.org/10.1080/00933104.2007.10473328

Memişoğlu, H. (2014). Sosyal bilgiler öğretmenlerinin görüşlerine göre vatandaşlık eğitimi [The educatıon of cıtızemshıp ın regard to the opınıons of socıal scıences teachers the objectıve of the study]. *Electronic Turkish Studies, 9*(5), 1565-184. Retrieved from https://web.p.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=7b4fc340-023f-483d-802a-3afa0b98c051%40redis

Merey, Z. (2009). İlköğretim sosyal bilgiler öğretiminde vatandaşlık ve insan hakları eğitimi [Citizenship and human rights education in primary school social studies teaching]. In M. Safran (Eds.) *Sosyal bilgiler öğretimi [Social studies education],* (pp. 719-742). Pegem Academy Publication.

Merey, Z., Karatekin, K., & Kuş, Z. (2012). İlköğretimde vatandaşlık eğitimi: Karşılaştırmalı bir çalışma [Citizenship education in primary education: A comparative study]. *Gazi University Journal of Gazi Educational Faculty, 32*(3), 795-821. Retrieved from https://www.researchgate.net/publication/312630956_Ilkogretimde_vatandaslik_egitimi_karsilastirmali_kuramsal_bir_calisma

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook.* Sage.

Millî Eğitim Bakanlığı [Ministry of National Education). (2018). *Sosyal bilgiler öğretim programları [Social studies curriculum].* Retrieved from https://mufredat.meb.gov.tr/ProgramDetay.aspx?PID=354

Önal, G., Öztürk, C., & Kenan, S. (2017). İngiltere vatandaşlık eğitimi programı: Toplumsal temelleri, boyutları ve temel bileşenleri [The UK citizenship education programme: Social foundations, dimensions and essential components]. *Anadolu Journal of Educational Sciences International*, *7*(2), 373-395. Retrieved from https://dergipark.org.tr/tr/download/article-file/333495

Önal, G., Öztürk, C., & Kenan, S. (2018). Teachers' perspectives on citizenship education in England. *Education and Science, 43*(196), 243-256. https://doi.org/10.15390/eb.2018.7060

Pederson, P.V., & Cogan J.J. (2003). Civic education pedagogy and assessment in the Pacific Rim: Six cases. *International Journal of Social Education, 17*(1), 16-30. Retrieved from https://dialnet.unirioja.es/info/textonodisponible

Sönmez, V. (1994). *Program geliştirmede öğretmen el kitabı [Teacher's handbook in curriculum development]*. Anı Publication.

Şen, A. (2019). Vatandaşlık eğitiminde değişiklik ve süreklilikler: 2018 sosyal bilgiler öğretim programı nasıl bir vatandaşlık eğitimi öngörüyor? [Changes and continuities in citizenship education: What kind of citizenship education does the 2018 social studies programme of study envisage?]. *Journal of Qualitative Research in Education, 7*(1), 1-28. https://dergipark.org.tr/en/pub/enad/issue/43049/521304

Şentürk, M., Şimşek, U., Tıkman, F., & Yıldırım, E. (2017). Sosyal bilgiler ve sınıf eğitimi öğretmen adaylarının gözünden vatandaşlık eğitimi: Nitel bir çalışma [The views of pre-service social studies teachers and pre-service classroom teachers about citizenship education: A qualitative study]. *Dicle University Journal of Ziya Gökalp Educational Faculty,* (32), 913-925. https://doi.org/10.14582/DUZGEF.1871

Üstel, F. (2005). *Makbul Vatandaş"ın peşinde II. Meşrutiyet'ten bugüne vatandaşlık eğitimi [In pursuit of the acceptable citizen II. Citizenship education from the Second Constitutional era to the present]*. İletişim Publications.

Wilkins, C. (2003). Teachers and young citizens: Teacher talk about their role as social educators. *Westminster Studies in Education, 26*(1), 63-75. https://doi.org/10.1080/0140672030260106

Wood, J. (2010) Preferred futures: active citizenship, government and young people's voices. *Youth & Policy,* 105, 50-70. Retrieved from https://www.researchgate.net/publication/344800107_'Preferred_Futures'_Active_Citizenship_Government_and_Young_People's_Voices

Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri [Qualitative research methods in the social sciences]*. Seçkin Publication.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 114–128

https://doi.org/10.21449/ijate.1183622

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

The power and type I error of Wilcoxon-Mann-Whitney, Welch's *t*, and student's *t* tests for likert-type data

Ahmet Salih Simsek [1,*]

[1]Kırsehir Ahi Evran University, Department of Measurement and Evaluation in Education, Türkiye

ARTICLE HISTORY

Received: Oct. 03, 2022

Accepted: Mar. 05, 2023

Keywords:

Likert,
Wilcoxon-Mann-Whitney,
Student’s t,
Welch’s t,
Unequal group size.

Abstract: Likert-type item is the most popular response format for collecting data in social, educational, and psychological studies through scales or questionnaires. However, there is no consensus on whether parametric or non-parametric tests should be preferred when analyzing Likert-type data. This study examined the statistical power of parametric and non-parametric tests when each Likert-type item was analyzed independently in survey studies. The main purpose of the study is to examine the statistical power of Wilcoxon-Mann-Whitney, Welch's t, and Student's *t* tests for Likert-type data, which are pairwise comparison tests. For this purpose, a Monte Carlo simulation study was conducted. The statistical significance of the selected tests was examined under the conditions of sample size, group size ratio, and effect size. The results showed that the Wilcoxon-Mann-Whitney test was superior to its counterparts, especially for small samples and unequal group sizes. However, the Student's *t*-test for Likert-type data had similar statistical power to the Wilcoxon-Mann-Whitney test under conditions of equal group sizes when the sample size was 200 or more. Consistent with the empirical results, practical recommendations were provided for researchers on what to consider when collecting and analyzing Likert-type data.

1. INTRODUCTION

Likert-type item is the most preferred response format for measuring characteristics of individuals in self-report instruments such as questionnaires and scales. The results of parametric or non-parametric analyses using this item type have been reported in many studies (Schrum *et al.,* 2020). One opinion in the literature is that Likert data can be considered interval data (Norman, 2010), while another opinion is that they are ordinal (Calver & Fletcher, 2020; Carifio & Perla, 2008). However, Likert scale data (i.e., the data that are the sum or mean of Likert items) can be analyzed under the assumption that they are interval data (Boone & Boone, 2017). For interval data, it is well known that the Student's *t*-test has better statistical power than the U-test in many cases when comparing means (Boneau, 1962; Glass et.al., 1972; Zimmerman & Zumbo, 1990; Bindak, 2014). Some researchers claim that this approach is also valid for data such as the Likert scale (Norman, 2010). Discussions on this issue can only be clarified with simulation studies.

*CONTACT: Ahmet Salih Şimşek ✉ asalihsimsek@gmail.com Kirsehir Ahi Evran University, Department of Measurement and Evaluation in Education, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

The results of comparative tests performed separately for Likert items are presented in many studies. Although non-parametric tests (e.g., Wilcoxon-Mann-Whitney) should be used for each Likert item considering its measurement level, it is observed that parametric tests (e.g., Student's t) are used (Liddel & Kruschke, 2018; Schrum *et al.,* 2020; Wu & Leung, 2017). It is clear that there is still confusion about which test should be preferred for Likert items. There are few simulation studies on which test is better than its counterpart for comparing groups (de Winter & Dodou, 2010; Derrick & White, 2017). This gap in knowledge in the literature leads to debates about the analysis of Likert data.

A simulation study reported that the Student's *t*-test and the U-test generally have similar power for five-point Likert items and that strong differences in power between the Student's *t*-test and the U-test occur when one of the samples is drawn from a multimodal distribution (de Winter & Dodou, 2010). The difference in power between the two tests for the fourteen different response patterns is also presented in the same study. In another study, Derrick & White (2017) examined the power of comparison tests for paired Likert data. The study, which examined the power of tests considering sample size, the correlation between paired observations, and the distribution of responses, emphasized that the paired samples *t*-test was not appropriate for paired Likert data (Derrick & White, 2017). The conditions considered in each simulation study differ from each other. Distributional characteristics (skewed, multimodal, normal, nonnormal, homogeneous, etc.), group size (equal, not equal), and sample size (small, medium, large) are the simulation conditions commonly used for between-group comparison tests. Testing all conditions in a single simulation run makes the interpretation of results difficult. For this reason, the performance of statistical tests is usually compared under certain selected simulation conditions.

The sample size is one of the parameters needed to calculate the statistical power of a test. For parametric mean comparison tests, it is desirable that the sample size for each group is not less than 30. However, previous studies have reported that Welch's *t*-test performs better than Student's *t*-test even for extremely small sample sizes (e.g., 2, 3, 5) (de Winter, 2013). On the other hand, a recent study highlighted that the desired statistical power cannot be achieved with the Student's *t*-test for small sample sizes (e.g., 15, 30, 50), so the sample size should not be less than 100 (Sangthong, 2020). Further simulation studies are needed on this topic.

Another important parameter affecting statistical power is the ratio of group sizes. It is well known that, especially for parametric tests, maximum power is achieved for a given total sample size when the groups are of equal size (1:1) (Kim & Park, 2019). Minimum required sample size to achieve the same level of power increases when group sizes are not equal (Bulus, 2021). The Student's *t* test is robust when the groups are equal in size, but in practice, almost all studies contain unequal group sizes (Ruxton, 2006). In another study, Ahad and Yahaya (2014) showed that the statistical power of Welch's test decreases dramatically under conditions of unequal group size. In the case of unequal group size (e.g., 1:2, 1:3, 1:4), the desired statistical power may not be achieved.

The relationship between effect size and the statistical power of the test is one of the most frequently asked questions in research (Bulus, 2021; Bulus, 2022; Bulus & Dong, 2021; Dong & Maynard, 2013). In addition, the statistical power of tests that compare the means of independent groups is not independent of effect size (Wiedermann & Eye, 2013). A higher effect size leads to higher statistical power (Ahad & Yahaya, 2014; de Winter, 2013). However, power differences between statistical tests often become apparent in cases where a smaller effect size is obtained. Therefore, it is necessary to evaluate the power of statistical tests under different effect size conditions.

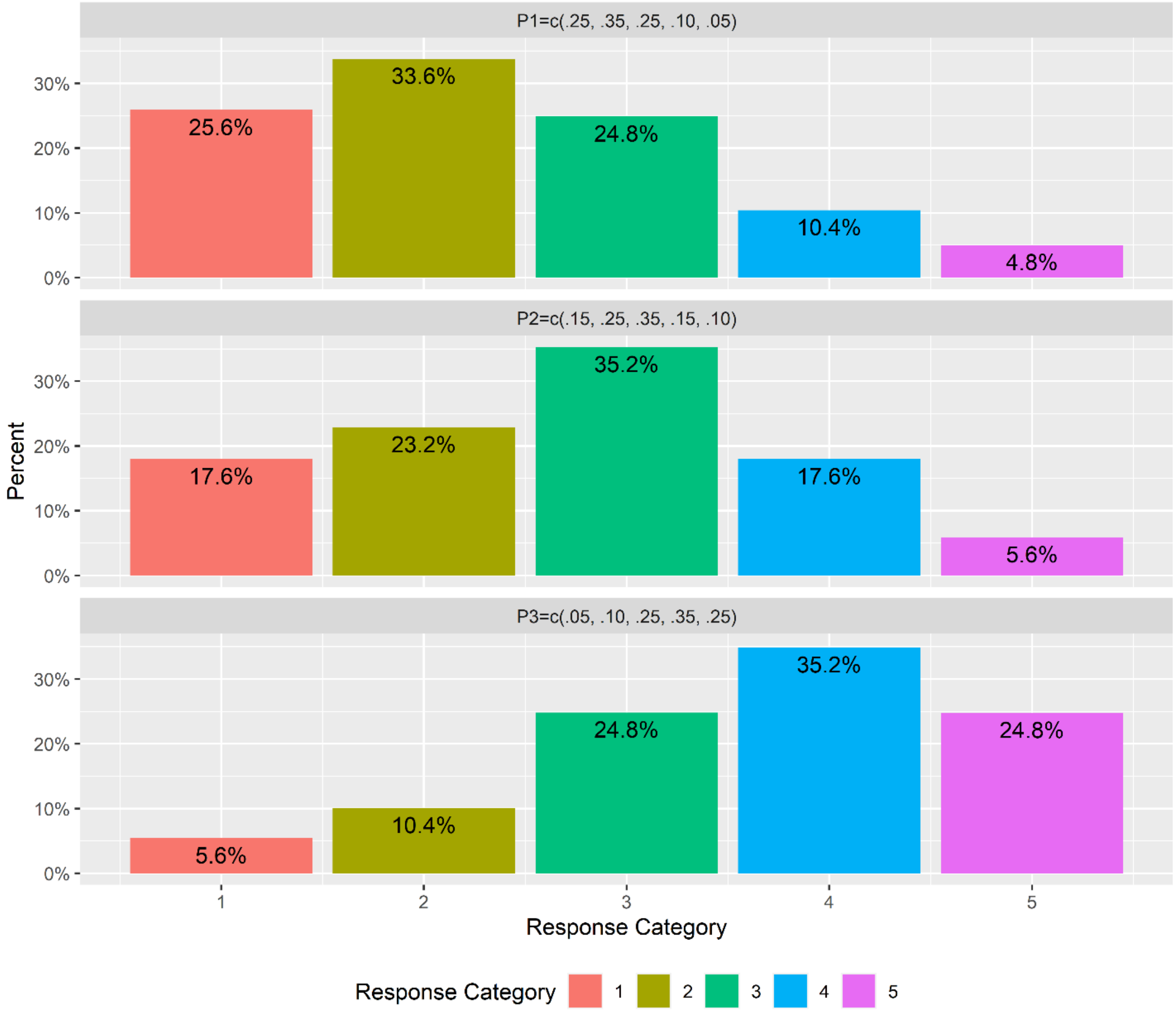
The current literature on the statistical power of tests comparing means between two groups is reviewed in this study. Although there is some disagreement about the effectiveness of parametric and nonparametric tests under different conditions, there is a lack of research on the use of parametric tests to compare the means of two groups for Likert scale data. Few studies have addressed this issue (de Winter & Dodou, 2010; Derrick & White, 2017). The gap in the literature regarding the lack of empirical studies examining different sample sizes, group size ratios, and effect sizes for Likert-type data is significant. This study is unique in that it aims to examine experimental conditions that have not previously been examined in selected comparison tests. Empirical evidence of the performance of parametric and nonparametric tests in analyzing Likert-type data under various conditions, such as sample size, group size ratio, and effect size, can help reduce uncertainties in the literature. Therefore, the objective of this study is to investigate the performance of Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for Likert-type data using the Monte Carlo simulation method under different sample characteristics. The study evaluates the statistical power of the selected parametric and nonparametric tests under simulated conditions with different sample sizes, group size ratios, and effect sizes. Based on the research findings, this study provides some guidelines for researchers interested in analyzing Likert-type data.

2. METHOD

The Monte Carlo simulation study was conducted with the data generated for a 5-point Likert item. During data generation, three different populations of respondents were designed in which the responses "disagree," "neutral," and "agree" predominated. The probability distribution of the responses designed for the three respondent populations according to the categories from 1 to 5 is P1=c(.25, .35, .25, .10, .05), P2=c(.15, .25, .35, .15, .10), and P3=c(.05, .10, .25, .35, .25), respectively. The probabilities determined for the response patterns are consistent with research outcomes in the literature (see de Winter & Dodou, 2010). The distribution of the categories of the generated data according to the population is shown in Figure 1. The data were generated with a total of 15000 observations, 5000 for each group.

In order to examine the statistical power of the selected tests at different effect sizes, samples were drawn from the P1-P2, P2-P3, and P1-P3 population pairs, and the statistical power of the selected tests was examined. In this way, the statistical power for the selected tests was calculated in the samples of populations with small ($.20<d\leq.50$), medium ($.50<d\leq1.00$), and large ($d>1.00$) effect sizes. The responses for populations P1, P2, and P3 were distributed with a mean of 2.34, 2.71, and 3.63, respectively. The effect size of the difference between populations in the generated data was determined to be 0.32 for P1-P2, 0.82 for P2-P3, and 1.15 for P1-P3. With three different effect sizes (small, medium, large), five different sample sizes (N=30, N=50, N=100, N=200, N=400), and three different group sizes ratios (1:1, 1:2, 1:4), the analyzes were performed with 5000 replicates for each condition. Table 1 shows the group sizes for the sample size and ratio of group size conditions. The results for the selected tests were obtained by analyzing the results of 225000 samples.

Figure 1. *Distribution of the response by Likert response category.*



Note. 5-point Likert response categories were expressed as 1 - disagree, 3 - neutral, and 5 - agree

Table 1. *The sample sizes and group sizes in the simulations*

	Sample size									
	N=30		N=50		N=100		N=200		N=400	
Group size ratio	n_1	n_2	n_1	n_2	n_1	n_2	n_1	n_2	n_1	n_2
1:1	15	15	25	25	50	50	100	100	200	200
1:2	10	20	17	34	33	67	66	134	133	267
1:4	6	24	10	40	20	80	40	160	80	320

Note. n_1 and n_2 indicate the group size for independent samples

Understanding the distributional properties of the data generated is a critical factor to consider when evaluating the results of comparative tests. With this in mind, the assumption of homogeneity of variance, a crucial assumption for parametric tests such as Student's *t*-test (Field, 2009), has been reviewed. The assumption of homogeneity of variances for the 225.000 samples drawn was tested using Levene's test. It was found that the assumption of homogeneity of variances was met for the majority of the samples (96% of all replications). In those cases where the homogeneity assumption was not met, a chi-square analysis was performed to identify if this was due to sample size and group size ratio bias. The results showed that there was no bias in either sample size (chi-square=0.2597, p=.878) or group size ratio (chi-square= 4.0562, p=.398). Likert-type data, by their nature, are interval data and are not expected to be normally distributed. However, Student's *t*-test, which is known to be robust to violations of the normality assumption (Bridge & Sawilowsky, 1999; Zimmerman, 1985). Heeren and D'Agostino (1999), show the robustness of the two independent samples *t*-tests to type I errors for Likert-type data when the sample size is small. One of the main objectives of the study is to investigate the statistical power and type I error rate of Student's *t*-test when the required assumptions are not met. As in similar simulation studies, the normality assumption for Likert-type data was not considered in this study. Data generation and analysis were performed using R software. Part of the simulation study codes can be found in the Appendix. Statistical power calculations were performed using the WMWssp (v.0.4.0) and MESS (v.0.5.7) packages. The statistical power of the Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for the simulation conditions were compared using statistical and graphical methods.

3. RESULTS

The study first examined the statistical power and type I error rate of the selected tests based on sample size and group size ratio. The mean values of the type I error rate and statistical power of the Wilcoxon-Mann-Whitney (WMW), Student's t, and Welch's *t*-tests for all effect size conditions were presented in Table 2 without being reported separately by effect size classification. The selected tests were evaluated by comparison with reference values, using an alpha level of 0.05 for the two-way hypothesis and a minimum statistical power of 0.80 suggested by Cohen (1988).

Table 2. *The type I error and the power of Wilcoxon-Mann-Whitney, Student's t, and Welch's t by sample size and group size ratio.*

Sample size	Group size ratio	Method	Power				Type I Error			
			Mean	sd	se	ci	Mean	sd	se	ci
N=30	1:1	WMW	0.540	0.33	0.003	0.005	0.051	0.22	0.003	0.006
N=30	1:1	Student's t	0.532	0.34	0.003	0.005	0.052	0.22	0.003	0.006
N=30	1:1	Welch's t	0.532	0.34	0.003	0.005	0.051	0.22	0.003	0.006
N=30	1:2	WMW	0.541	0.34	0.003	0.005	0.047	0.21	0.003	0.006
N=30	1:2	Student's t	0.505	0.33	0.003	0.005	0.052	0.22	0.003	0.006
N=30	1:2	Welch's t	0.496	0.33	0.003	0.005	0.051	0.22	0.003	0.006
N=30	1:4	WMW	0.550	0.34	0.003	0.005	0.044	0.21	0.003	0.006
N=30	1:4	Student's t	0.435	0.31	0.003	0.005	0.048	0.21	0.003	0.006
N=30	1:4	Welch's t	0.392	0.29	0.002	0.005	0.056	0.23	0.003	0.006
N=50	1:1	WMW	0.654	0.34	0.003	0.005	0.050	0.22	0.003	0.006
N=50	1:1	Student's t	0.645	0.34	0.003	0.005	0.050	0.22	0.003	0.006
N=50	1:1	Welch's t	0.645	0.34	0.003	0.005	0.050	0.22	0.003	0.006
N=50	1:2	WMW	0.658	0.34	0.003	0.005	0.048	0.21	0.003	0.006
N=50	1:2	Student's t	0.614	0.34	0.003	0.005	0.049	0.22	0.003	0.006
N=50	1:2	Welch's t	0.609	0.34	0.003	0.005	0.052	0.22	0.003	0.006
N=50	1:4	WMW	0.657	0.34	0.003	0.005	0.052	0.22	0.003	0.006
N=50	1:4	Student's t	0.551	0.34	0.003	0.005	0.053	0.22	0.003	0.006
N=50	1:4	Welch's t	0.526	0.33	0.003	0.005	0.054	0.23	0.003	0.006
N=100	1:1	WMW	0.784	0.31	0.003	0.005	0.051	0.22	0.003	0.006
N=100	1:1	Student's t	0.773	0.32	0.003	0.005	0.050	0.22	0.003	0.006
N=100	1:1	Welch's t	0.773	0.32	0.003	0.005	0.050	0.22	0.003	0.006
N=100	1:2	WMW	0.787	0.31	0.003	0.005	0.048	0.21	0.003	0.006

N=100	1:2	Student's t	0.758	0.32	0.003	0.005	0.050	0.22	0.003	0.006
N=100	1:2	Welch's t	0.756	0.32	0.003	0.005	0.052	0.22	0.003	0.006
N=100	1:4	WMW	0.782	0.31	0.003	0.005	0.046	0.21	0.003	0.006
N=100	1:4	Student's t	0.702	0.34	0.003	0.005	0.048	0.21	0.003	0.006
N=100	1:4	Welch's t	0.693	0.34	0.003	0.005	0.051	0.22	0.003	0.006
N=200	1:1	WMW	0.874	0.24	0.002	0.004	0.052	0.22	0.003	0.006
N=200	1:1	Student's t	0.860	0.25	0.002	0.004	0.052	0.22	0.003	0.006
N=200	1:1	Welch's t	0.860	0.25	0.002	0.004	0.052	0.22	0.003	0.006
N=200	1:2	WMW	0.873	0.24	0.002	0.004	0.048	0.21	0.003	0.006
N=200	1:2	Student's t	0.847	0.27	0.002	0.004	0.048	0.21	0.003	0.006
N=200	1:2	Welch's t	0.846	0.27	0.002	0.004	0.049	0.22	0.003	0.006
N=200	1:4	WMW	0.867	0.26	0.002	0.004	0.051	0.22	0.003	0.006
N=200	1:4	Student's t	0.808	0.30	0.002	0.005	0.053	0.22	0.003	0.006
N=200	1:4	Welch's t	0.805	0.30	0.002	0.005	0.055	0.23	0.003	0.006
N=400	1:1	WMW	0.950	0.13	0.001	0.002	0.045	0.21	0.003	0.006
N=400	1:1	Student's t	0.937	0.15	0.001	0.002	0.049	0.22	0.003	0.006
N=400	1:1	Welch's t	0.937	0.15	0.001	0.002	0.049	0.22	0.003	0.006
N=400	1:2	WMW	0.947	0.14	0.001	0.002	0.045	0.21	0.003	0.006
N=400	1:2	Student's t	0.924	0.17	0.001	0.003	0.045	0.21	0.003	0.006
N=400	1:2	Welch's t	0.923	0.17	0.001	0.003	0.047	0.21	0.003	0.006
N=400	1:4	WMW	0.940	0.16	0.001	0.003	0.052	0.22	0.003	0.006
N=400	1:4	Student's t	0.888	0.22	0.002	0.004	0.052	0.22	0.003	0.006
N=400	1:4	Welch's t	0.887	0.22	0.002	0.004	0.051	0.22	0.003	0.006

Note: se: Standard error; ci: %95 confidence interval

The results for sample size show that statistical power of .80 and above is achieved for samples of N=200 and above. Regardless of the method chosen, statistical power increases in parallel with the increase in sample size. The Wilcoxon-Mann-Whitney method provides higher statistical power than the other methods, and the difference in power increases for smaller samples. Student's t and Welch's *t*-tests showed similar power for the sample size condition. The results of the group size ratio show that the Wilcoxon-Mann-Whitney method has a relatively small advantage over the other methods for the same group size, while the Wilcoxon-Mann-Whitney method has a significant power advantage over the other methods as the difference between group sizes increases. Depending on the group size ratio, higher statistical power was obtained for Wilcoxon-Mann-Whitney, Student's t, and Welch's t.

When evaluating the Type I error rate, a lower Type I error was found for the Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests when the sample size was 400. In this context, our analyses have revealed that the sample size leads to a lower Type I error rate than expected. Compared to a group size ratio of 1:4, the type I error rate has proved to be more conservative and stable under the conditions of 1:1 and 1:2. The results show that in the case where the group size ratio is the highest, the type I error rate is not consistent, and the type I error rate may exceed the expected value of 5% depending on the sample size. The results indicate that the sample size and group size ratio for Likert data have a lower Type I error than the Wilcoxon-Mann-Whitney test, Student's *t*-test, and Welch's test.

Figure 2 and Figure 3 indicate how the empirical power and Type I error rate of the Wilcoxon-Mann-Whitney, Student's t, and Welch's t methods change as a function of the interaction between sample size and group size ratio.

Figure 2. *Empirical power for Wilcoxon-Mann-Whitney, Student's t, and Welch's t in terms of sample size and group size ratio*

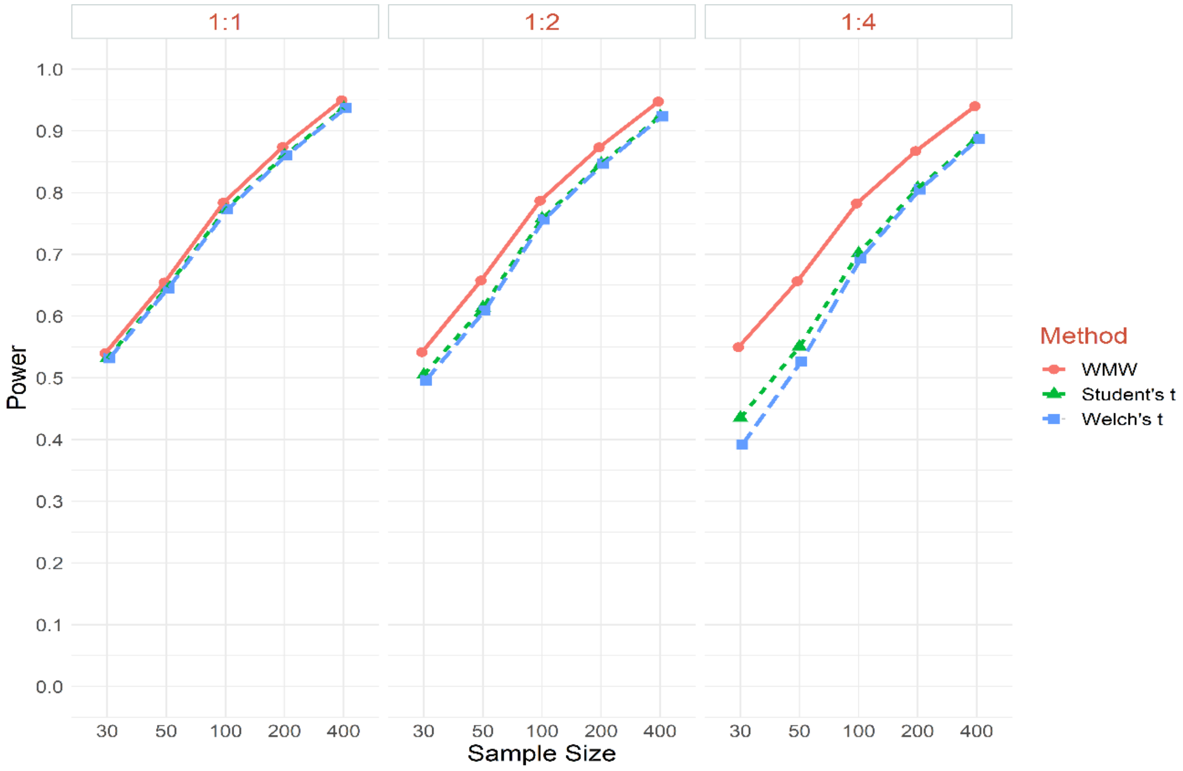


Figure 3. *Type I error rates for Wilcoxon-Mann-Whitney, Student's t, and Welch's t in terms of sample size and group size ratio*

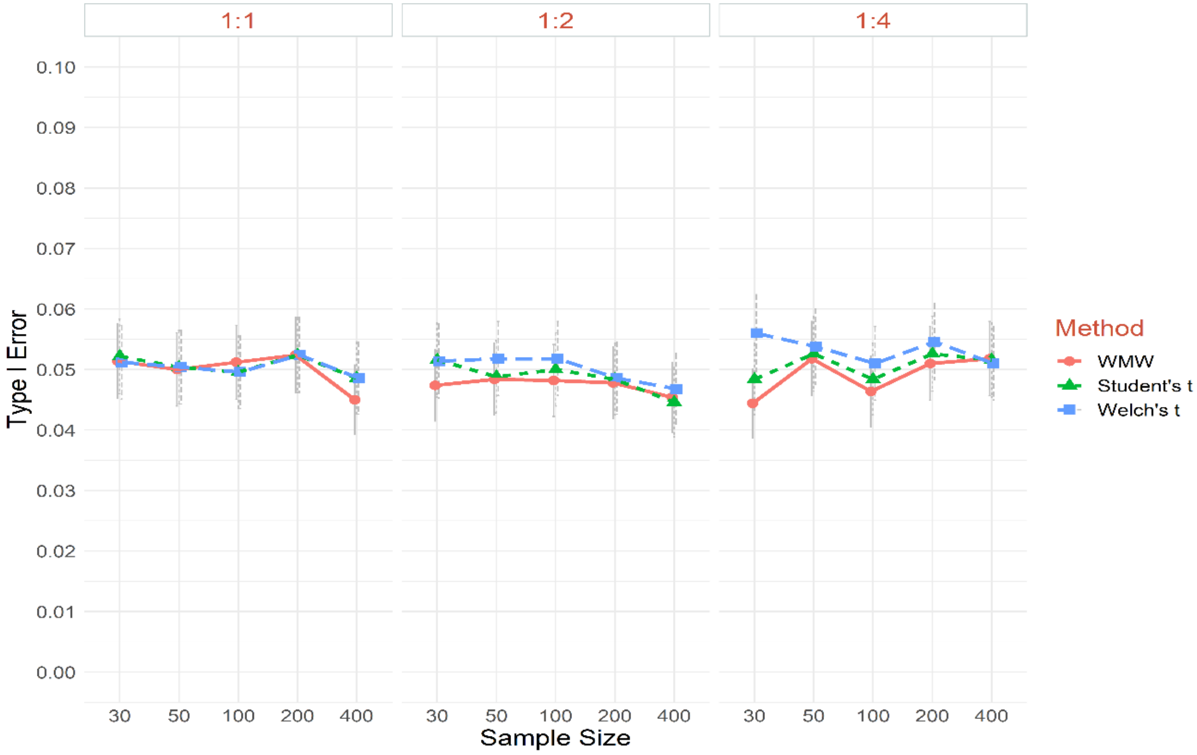
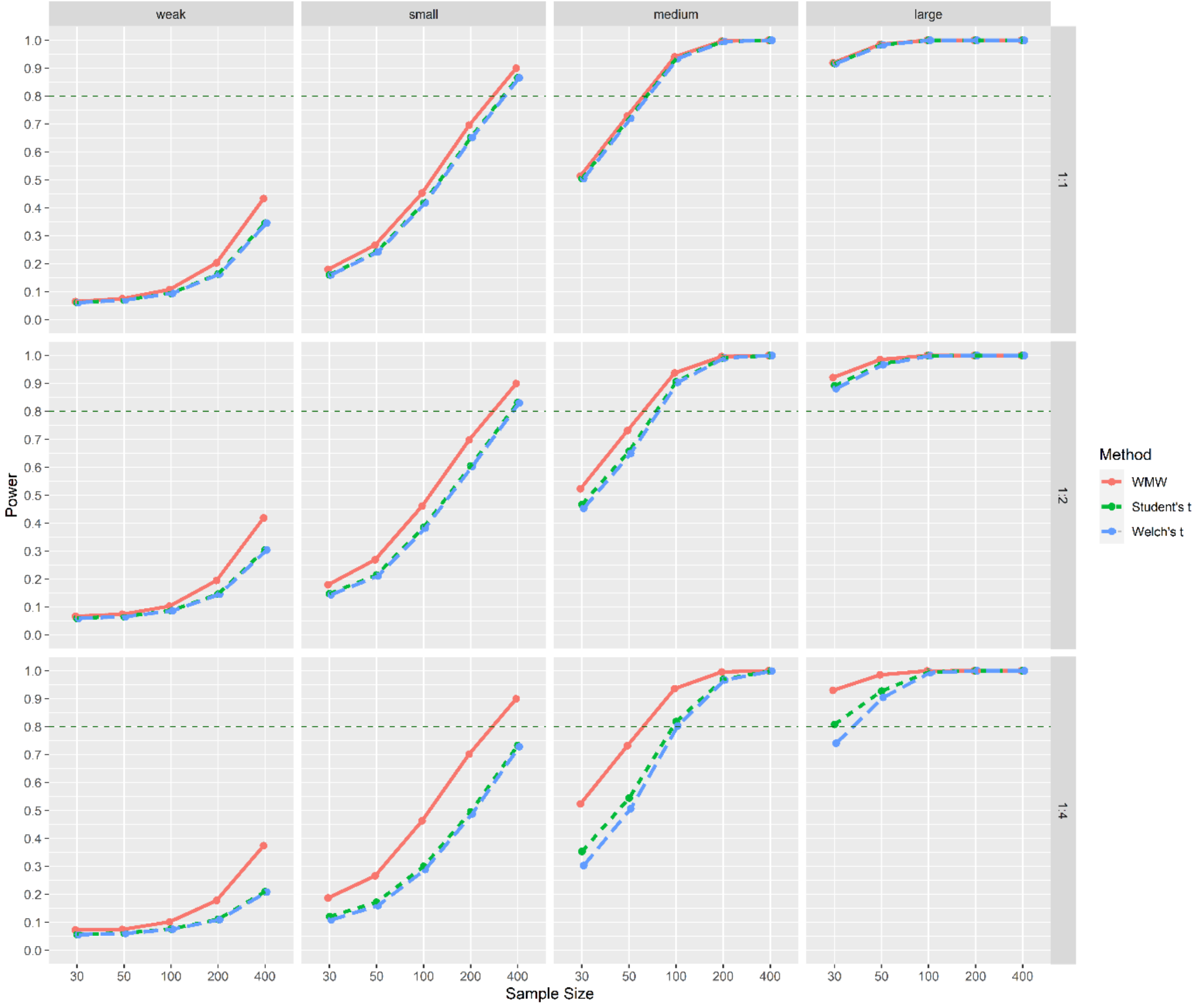


Figure 2 indicates that Wilcoxon-Mann-Whitney outperforms the other methods in terms of both sample size and group size ratio. In particular, when group sizes are unequal, Wilcoxon-Mann-Whitney seems to offer a more significant performance advantage than other methods.

Figure 4 indicates how the Wilcoxon-Mann-Whitney, Student's t, and Welch's t power functions change for effect size, sample size, and group size ratio. Looking at the line plots for the statistical power functions at weak and small effect sizes, the statistical power for the selected tests increases as the sample size increases and the group size ratio approaches 1:1. Cohen (1988) suggests that at least .80 should be used as a threshold for statistical power. For this reason, statistical power of .80 was accepted as the lower limit for evaluating the performance of the compared tests. However, it was found that the statistical power of .80 could not be achieved under any weak effect size condition, while it was achieved at N=400 for the small effect size condition. Moreover, the advantage of the power function of Wilcoxon-Mann-Whitney came into play when the difference between weak and small effect sizes, sample size, and group sizes increased. Under the conditions of medium and large effect size, Wilcoxon-Mann-Whitney responded more conservatively to changes in both sample size and group size ratio, while other methods were affected by the changes. In particular, under the condition that the group size ratio was 1:4 and the sample size was 30, Wilcoxon-Mann-Whitney had a remarkable advantage over other methods. However, it was found that although the group size ratio varied with a large effect size for samples of 50 and above, other methods offered statistical power that was close to Wilcoxon-Mann-Whitney.

Figure 4. *Line plot of the power function for effect size x sample size x group size ratio.*



4. DISCUSSION and CONCLUSION

Survey studies using Likert-type items are very common in the social sciences, e.g., education, psychology, and health. When each Likert item must be analyzed independently, researchers are faced with the question of which parametric or nonparametric tests to use. This study examined the performance of Student's t, Welch's t, and Wilcoxon-Mann-Whitney tests in analyzing Likert items for two independent groups. Under all conditions examined in the simulation studies, the Wilcoxon-Mann-Whitney test performed similarly well or better than its counterparts, Student's t and Welch's *t*-tests. The Type I error rate was found to be lower for Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests when the sample size was 400. The results revealed that a lower Type I error rate was achieved as the sample size increased, and the type I error rate was more conservative and stable under the conditions of 1:1 and 1:2 group size ratios. On the other hand, when the group size ratio was the highest, the type I error rate was found to be inconsistent and may exceed the expected value of 5% depending on the sample size. One reason why the Mann-Whitney U test may be a better choice for analyzing Likert-type data is that it measures the median difference between two groups, as opposed to the mean difference measured by the Student's *t*-test. In addition, the Mann-Whitney U test is also known to be more informative than Student's *t*-test when the modal value deviates from the mean, it is less sensitive to outliers, and it does not require a normality assumption (Wilcox, 2012; Field, 2009).

In contrast to previous studies, this study examined the statistical power of the selected tests in terms of sample size, group size ratio, and effect size for Likert-item analysis. The results obtained for sample size show that the statistical power of all tests is affected by sample size. These results are similar to those obtained in previous studies for continuous data (Dwivedi *et al.,* 2017; Ma *et al.,* 2021; Sangthong, 2020; Wiedermann & Eye, 2013). The results suggest that the findings on sample size for continuous data also apply to Likert-type data. The results also showed that Wilcoxon-Mann-Whitney was stronger than its counterparts under all sample size conditions. Previous studies support these empirical findings (de Winter & Dodou, 2010; Nanna & Sawilowsky, 1998). The second major finding of the study relates to the group size ratio. The results show that Student's t and Welch's t are sensitive to changes in the group size ratio. When the group size ratio is 1:1 for both tests, the statistical power is similar to Wilcoxon-Mann-Whitney, but when the ratio changes, the statistical power decreases. Wilcoxon-Mann-Whitney provides better power for Likert data than its counterparts when the group size ratio changes. These results are confirmed by some studies in the literature (Ahad & Yahaya, 2014; de Winter & Dodou, 2010; Dwivedi *et al.,* 2017; Zimmerman, 2004).

When the statistical power of the tests was examined considering the interaction of sample size and group size ratio, Student's t and Welch's t achieved the desired statistical power when the sample size was 200 or more. Likert-type data are defined as ordinal or interval values. Since the data structure is suitable for the structure of Wilcoxon-Mann-Whitney, it can be stated that it performs better than its counterparts. However, an important finding of the study is that Student's t and Welch's *t*-tests perform similarly well as Wilcoxon-Mann-Whitney for large samples (N > 200) and the same group ratio (1:1). This result implies that parametric tests such as Student's t can be used for the analysis of Likert-type data under certain conditions. Finally, the statistical power of the tests selected by effect size showed that high statistical power was obtained even for small samples for medium and large effect sizes. However, for the weak effect size, the desired statistical power was not achieved even with a sample size of 400. For the small effect size, it was found that the selected statistical tests could achieve the desired statistical power with a sample size of 400.

4.1. Implications

This study examined the performance of pairwise comparison tests for independent groups under conditions of sample size, group size ratio, and effect size. Considering the research findings and limitations, some suggestions for future research and researchers were developed.

4.1.1. *Methodological implications*

The results of this study have significant methodological implications for the appropriate selection of statistical tests when analyzing data collected using Likert-type scales. The simulation studies demonstrated that the Wilcoxon-Mann-Whitney test was either comparable or superior to Student's t and Welch's t-tests in terms of Type I error rate and statistical power. Additionally, the Mann-Whitney U test emerged as a suitable alternative for the Student's *t*-test when analyzing Likert-type data, as it measures the median difference between two groups and is less sensitive to outliers, and does not require a normal distribution assumption. The study results also indicated that the statistical power of the selected tests was dependent on both sample size and the ratio of group sizes, and the Wilcoxon-Mann-Whitney test was found to be more robust under all conditions. However, it is worth noting that the simulation studies were based on a distribution commonly found in Likert items while actual response distributions may vary due to human subjectivity. To validate the findings, further research using real-world survey data with high participation rates (such as PISA and TIMSS) is recommended. The impact of missing data, which is frequently encountered in real-world applications, should also be explored in future studies.

4.1.2. *Practical implications*

The study provides practical recommendations for researchers using Likert-type data in their studies. Based on the results, the Wilcoxon-Mann-Whitney test is recommended as the preferred choice over Student's *t*-test and Welch's *t*-test when analyzing Likert-type data because it provides higher statistical power. In addition, the Student's t and Welch's *t*-tests are as robust as the Wilcoxon-Mann-Whitney test when the sample size is 200 or more and the group size is the same. However, when the sample size is less than 100 and the ratio of group sizes is unequal, the Wilcoxon-Mann-Whitney test should be preferred because it provides higher power. It is emphasized that sample size, group size ratio, and effect size should be considered when selecting a pairwise comparison test for Likert-type data to achieve the desired level of statistical power.

4.2. Limitations

The present study has some limitations regarding the generalizability of the results. Specifically, the results on statistical power and Type I error rates for the tests are limited to the specific conditions examined in this study. These conditions include the use of Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for Likert-type data in independent sample environments and sample sizes of 30, 50, 100, 200, and 400, which are commonly used in survey studies. However, the applicability of these results to smaller or larger sample sizes needs to be investigated in future studies. In addition, the power of the tests examined in this study is limited to group size ratios of 1:1, 1:2, and 1:4, which may not cover all possible group size ratios in survey studies. Therefore, further research is recommended to investigate the performance of these tests at different group sizes. Another limitation of this study is the use of only three response patterns (disagree, neutral, and agree) as a reference for the generated data. The scatter properties of the data were not considered when determining the effect size. Therefore, future studies should investigate the performance of the selected tests under different response patterns and distributional properties. In summary, although the present study provides valuable insights into the performance of Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for Likert data under certain conditions, the generalizability of the results is limited with the simulation conditions. In order to improve the external validity of these results, future studies should consider a wider range of sample sizes, group sizes, response patterns, and distributional characteristics.

Acknowledgments

Acknowledgments of people, grants, and funds should be placed in a separate section before the References. If the study has been previously presented at a conference or a scholarly meeting, it should be mentioned here.

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

Ahmet Salih Şimşek https://orcid.org/0000-0002-9764-3285

REFERENCES

Ahad, N.A., & Yahaya, S.S.S. (2014). Sensitivity analysis of Welch's *t*-test. AIP Conference Proceedings, 1605(February 2015), 888–893. https://doi.org/10.1063/1.4887707

Bindak, R. (2014). Comparison Mann-Whitney U Test and Students' *t* Test in Terms of Type I Error Rate and Test Power: A Monte Carlo Sımulation Study. *Afyon Kocatepe University Journal of Sciences and Engineering, 14*, 5-11. https://doi.org/10.5578/fmbd.7380

Boneau, C.A. (1962). A comparison of the power of the U and *t*-tests. Psychological Review, 69, 246-256. https://doi.org/10.1037/h0047269

Boone, H.N., Boone, D.A. 2012. Analyzing Likert data. *Journal of Extension, 50*(2), 1-5. Retrieved February 20, 2023, from https://eric.ed.gov/?id=EJ1042448

Bridge, P.D., & Sawilowsky, S.S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the *t*-test and Wilcoxon Rank-Sum test in small samples applied research. *Journal of clinical epidemiology, 52*(3), 229-35. https://doi.org/10.1016/S0895-4356(98)00168-1

Bulus, M. (2021). Sample size determination and optimal design of randomized/non-equivalent pretest-posttest control-group designs. *Adiyaman Univesity Journal of Educational Sciences, 11*(1), 48-69. https://doi.org/10.17984/adyuebd.941434

Bulus, M. (2022). Minimum detectable effect size computations for cluster-level regression discontinuity: Specifications beyond the linear functional form. *Journal of Research on Education Effectiveness, 15*(1), 151-177. https://doi.org/10.1080/19345747.2021.1947425

Bulus, M., & Dong, N. (2021). Bound-constrained optimization of sample sizes subject to monetary restrictions in planning multilevel randomized trials and regression discontinuity studies. *The Journal of Experimental Education, 89*(2), 379-401. https://doi.org/10.1080/00220973.2019.1636197

Calver, M., & Fletcher, D. (2020). When ANOVA isn't ideal: Analyzing ordinal data from practical work in biology. *The American Biology Teacher, 82*(5), 289-294. https://doi.org/10.1525/abt.2020.82.5.289

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical education, 42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Champagne, C.A., & Curran, P.J. (2017). Using Monte Carlo simulations to demonstrate the importance of statistical power. *The Journal of Educational Research, 110*(6), 524-532. https://doi.org/10.1080/00220671.2015.1079697

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

de Winter, J.F., & Dodou, D. (2010). Five-point Likert items: *t*-test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research, and Evaluation, 15*(1), 11. https://doi.org/10.7275/bj1p-ts64

de Winter, J.F. (2013) Using the Student's *t*-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18, 10. https://doi.org/10.7275/e4r6-dj05

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology, 30*(1), 92. https://www.rips-irsp.com/articles/10.5334/irsp.661/

Derrick, B., & White, P. (2017). Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics, 18*(3). Retrieved February 20, 2023, from http://www.ceser.in/ceserp/index.php/ijms/article/view/4997

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67. https://doi.org/10.1080/19345747.2012.673143

Dwivedi, A.K., Mallawaarachchi, I., & Alvarado, L.A. (2017). Analysis of small sample size studies using non-parametric bootstrap test with pooled sampling method. *Statistics in Medicine, 36*, 2187 - 2205. https://doi.org/10.1002/sim.7263

Field, A. (2009). Discovering statistics using SPSS (3rd ed.). Sage publications.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288. https://doi.org/10.3102/00346543042003237

Harpe, S.E. (2015). How to analyze Likert and other rating scale data. Currents in Pharmacy Teaching and Learning, 7, 836-850. https://doi.org/10.1016/j.cptl.2015.08.001

Heeren, T., & D'Agostino, R.B. (1987). Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Statistics in Medicine, 6*(1), 79-90. https://doi.org/10.1002/sim.4780060110

Jamieson S. (2004). Likert scales: how to (ab)use them. *Medical education, 38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Kim, T.K., & Park, J.H. (2019). More about the basic assumptions of *t*-test: normality and sample size. *Korean Journal of Anesthesiology, 72*(4), 331-335. https://doi.org/10.4097/kja.d.18.00292

Liddell, T.M., & Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology, 79*, 328-348. https://doi.org/10.1016/j.jesp.2018.08.009

Ma, C., Wang, X., Xia, L., Cheng, X., & Qiu, L. (2021). Effect of sample size and the traditional parametric, non-parametric, and robust methods on the establishment of reference intervals: Evidence from real-world data. *Clinical Biochemistry, 92*, 67–70. https://doi.org/10.1016/j.clinbiochem.2021.03.006

Nanna, M.J., & Sawilowsky, S.S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3*(1), 55-67. https://doi.org/10.1037/1082-989X.3.1.55

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Ruxton, G.D. (2006). The unequal variance Student's *t* testis an underused alternative to Student's *t* test and the Mann–Whitney U test. *Behavioral Ecology, 17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

Sangthong, M. (2020). The Effect of the Likert Point Scale and Sample Size on the Efficiency of Parametric and Non-parametric Tests. *Thailand Statistician, 18*(1), 55–64.

Schrum, M.L., Johnson, M., Ghuy, M., & Gombolay, M.C. (2020). *Four years in review: Statistical practices of Likert scales in human-robot interaction studies*. In Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 43-52). https://doi.org/10.1145/3371382.3380739

Wiedermann, W., & von Eye, A. (2013). Robustness and power of the parametric *t*-test and the non-parametric Wilcoxon test under non-independence of observations. *Psychological Test and Assessment Modeling, 55*(1), 39-61.

Wilcox, R.R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press.

Wu, H., & Leung, S.O. (2017). Can Likert scales be treated as interval scales? Simulation study. *Journal of Social Service Research, 43*(4), 527-532. https://doi.org/10.1080/01488376.2017.1329775

Zimmerman D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173-181. https://doi.org/10.1348/000711004849222

Zimmerman, D.W. & Zumbo, B.D. (1990) The Relative Power of the Wilcoxon-Mann-Whitney Test and Student *t* Test Under Simple Bounded Transformations. *The Journal of General Psychology, 117*(4), 425-436, https://doi.org/10.1080/00221309.1990.9921148

Zimmerman, D.W. (1985). Power Functions of the *t*-test and Mann-Whitney U Test Under Violation of Parametric Assumptions. *Perceptual and Motor Skills, 61*, 467 - 470. https://doi.org/10.2466/pms.1985.61.2.467

APPENDIX

```
for (i in 1:5) {  #SS
  for (j in 1:3) {  #Dist
    for (k in 1:100) {  #iterid
iterid<-k
Nsize<-Ssize[i]  #Nsizeal data size
N1ratio<-Sprop[j]  #Group distribution proportion
#-- GROUP STAT
n1<-round(Nsize*N1ratio,0)
n2<-Nsize-n1
nmin<-min(n1,n2)
Nratio<-round((1-N1ratio)/N1ratio,0)
sG1<-sample(dG1,n1)
sG2<-sample(dG2,n2)
Diff.p = abs(mean(sG1)-mean(sG2))/min(mean(sG1),mean(sG2))
#-- cohen's d effect size
SD1<-(n1-1)*((sdG1)^2)
SD2<-(n2-1)*((sdG2)^2)
sdpooled<-sqrt((SD1+SD2)/(n1+n2-2))
d<-abs((mG1-mG2)/sdpooled)
#--- POWER TEST for WMW
#library(WMWssp)
p_WMW<-as.numeric(WMWssp_maximize(sG1, sG2, alpha = alpha, N=Nsize)$power)
#--- POWER TEST for t-test
#library("MESS") -> unequal group size
p_ttest<-as.numeric(power_t_test(n=nmin, delta=d, sig.level = alpha, ratio=Nratio, sd.ratio=SDratio, df.method ="classical", type = "two.sample", alternative="two.sided")$power)
p_welch<-as.numeric(power_t_test(n=nmin, delta=d, sig.level = alpha, ratio=Nratio, sd.ratio=SDratio, df.method ="welch", type = "two.sample", alternative="two.sided")$power)
#--- p value for WMW
#library(rcompanion)
WMW.z<-as.numeric(wilcoxonZ(sG1,sG2))
WMW.d<-as.numeric(wilcoxonRG(WMW_Y,WMW_G))
WMW.w<-as.numeric(wilcox.test(sG1,sG2, alternative = "two.sided", exact = FALSE)$statistic)
WMW.p<-as.numeric(wilcox.test(sG1,sG2, alternative = "two.sided", exact = FALSE)$p.value)
#--- p value for t-test
ttest.t<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = TRUE)$statistic)
ttest.p<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = TRUE)$p.value)
#--- p value for welch
welch.t<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = FALSE)$statistic)
welch.p<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = FALSE)$p.value)
#--- Homogeniy of Variance Analysis
#library(DescTools)
WMW_Y<-c(sG1,sG2)
WMW_G<-factor(c(rep("1", length(sG1)),  rep("2", length(sG2))))
levene.p<-as.numeric(LeveneTest(WMW_Y,WMW_G)$`Pr(>F)`[1])
levene.factor<-ifelse(levene.p>.05, 1, 0)
#--- Normality
#library(DescTools)
G1.p.norm<-shapiro.test(sG1)$p.value
G2.p.norm<-shapiro.test(sG2)$p.value
G1.factor.norm<-ifelse(G1.p.norm>.05, 1, 0)
G2.factor.norm<-ifelse(G2.p.norm>.05, 1, 0)
interimdata<-as.vector(cbind(simid,iterid,iNsize,iNratio,
                    n1,n2,d,d.cat,
                    p_Wilcoxon-Mann-Whitney,p_ttest,p_welch,
                    mG1,mG2,sdG1,sdG2,
                    WMW.z,WMW.d,
                    WMW.w,WMW.p,
                    ttest.t,ttest.p,
                    welch.t,welch.p,
                    levene.p,levene.factor,
                    G1.p.norm,G2.p.norm,
                    G1.factor.norm,G2.factor.norm))
simdata<-rbind(simdata,interimdata)
simid<-simid+1
iterid<-k+1
} j<-j+1} i<-i+1}
```

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 129–144

https://doi.org/10.21449/ijate.1148460

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

Modeling unobserved heterogeneity using person-centered approaches: Latent profiles of preservice teachers' emotional awareness

Esra Sozer-Boz[1,*], Derya Akbas[2], Nilufer Kahraman[3]

[1]Bartin University, Faculty of Education, Department of Educational Sciences, Türkiye
[2]Aydın Adnan Menderes University, Faculty of Education, Department of Educational Sciences, Türkiye
[3]Gazi University, Faculty of Gazi Education, Department of Educational Sciences, Türkiye

ARTICLE HISTORY

Received: July 25, 2022

Accepted: Mar. 08, 2023

Keywords:

Latent profile,
Heterogeneity,
Preservice teachers,
Emotional awareness.

Abstract: Latent Class and Latent Profile Models are widely used in psychological assessment settings, especially when individual differences are suspected to be related to unobserved class memberships, such as different personality types. This paper provides an easy-to-follow introduction and application of the methodology to the data collected as part of more extensive educational research investigating social-emotional competency profiles of preservice teachers (n=184) who responded to an Emotional Awareness Questionnaire. Suspected that there would be two or more latent emotional awareness sub-groups in the sample, a series of latent profile models was estimated. The results suggested three distinct emotional awareness profiles; namely, introverted, extroverted, and less sensitive to others' emotions, with proportions of 9%, 56%, and 35%, respectively. Subsequent analyses showed that preservice teachers with higher levels of emotionality, sociability, and well-being were more likely to be in the extroverted profile. The findings suggest that nearly half of the teachers in the sample could be expected to possess the most professionally desirable teacher profile. Nonetheless, it was noted that if timely diagnostic and tailored training or intervention programs were available, at least some of the preservice teachers in the less sensitive to others' profiles, and most of the preservice teachers in the introverted profile could be helped to self-observe the way which they tend to identify and regulate their emotions.

1. INTRODUCTION

Individual differences represent an important issue for educators and researchers (Snow, 1986) since individuals of any age and culture differ in various cognitive, affective, and psychomotor skills, which are directly related to differences in individuals' learning and growth processes. To this end, many kinds of research strive to characterize patterns and pathways of individuals' development (Hickendorff et al., 2018). Furthermore, development can occur in stages, growth patterns can vary between individuals, and growth can interact with individuals' characteristics. The development of populations in educational and psychological sciences is often heterogeneous, while population heterogeneity can be observed or unobserved. Heterogeneity

*CONTACT: Esra Sozer-Boz ✉ esrsozer@gmail.com Bartin University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

is observed if it is possible to define the subpopulations based on an observed variable such as gender, control, and experimental groups. In observed heterogeneity, group membership is known for each participant. However, the sources of unobserved heterogeneity may not be known a priori (Lubke & Muthen, 2005) and disregarding the unobserved heterogeneity in investigating individual differences may cause inadequate descriptions for many individuals in a population (Hickendorff et al., 2018). When the subpopulation membership of the participants is not observed, group memberships should be inferred from the data collected. In the context of unobserved heterogeneity, subpopulations are called latent classes or profiles. Therefore, assessing and modeling the heterogeneity is essential for understanding how and under which circumstances growth occurs. In such cases, the researcher may use the latent profile or latent class analyses to model the unobserved heterogeneity between and within individuals more appropriately.

In social, behavioral, and educational sciences, programs are often administered to populations without consideration of individual characteristics. Recently, there has been growing interest in individualizing treatments to administer the right program to the right individuals to maximize the effectiveness of such treatments (Lanza & Rhoades, 2013). In this context, person-centered approaches have become more helpful in investigating unobserved heterogeneity in a population (Jung & Wickrama, 2008). Latent class analysis (LCA) and latent profile analysis (LPA) are person-centered approaches tracing back heterogeneity in a population to some existing but unobserved sub-groups of individuals (Hickendorff et al., 2018). LPA identifies heterogeneity in cross-sectional data by grouping participants into latent classes based on similarities in the continuous observed/indicator variables.

LCA and LPA are in the Finite Mixture Modeling framework (Gibson, 1959; Hickendorff et al., 2018; Peugh & Fan, 2013), referring to a class of statistical analysis techniques designed to model unobserved population heterogeneity by grouping individuals. Mixture models have different names depending on whether the observed and latent variables are continuous or categorical. These models are shown in Table 1, in which the rows correspond to continuous and categorical observed variables and the columns to continuous and categorical latent variables. LPA determines latent groups using continuous observed variables, and LCA does the same using categorical variables (Oberski, 2016).

Table 1. *Latent variable models*[*]

		Latent Variables	
		Categorical	Continuous
Observed Variables	Categorical	Latent Class Analysis	Item Response Theory
	Continuous	Latent Profile Analysis	Factor Analysis

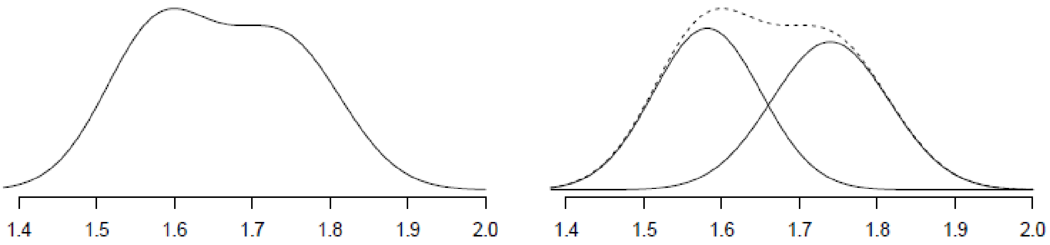
*Muthén, B. (2007). Latent variable hybrids: Overview of old and new methods. In G.R. Hancock & K.M. Samuelsen (Eds.), *Advances in latent variable mixture modeling* (pp. 1-24). Information Age.

LPA models are similar to clustering methods, while they have a more flexible structure (Tein et al., 2013). The primary goal of LPA is to maximize the homogeneity within groups (i.e., individuals within a profile should be similar) and maximize the heterogeneity between groups (i.e., individuals between profiles should be different). These groups are represented by a categorical latent variable, as they are not directly known but inferred from observed variables' response patterns (Roesch et al., 2010). Identifying latent profiles can be useful for characterizing qualitative and quantitative inter-and intra- individual differences simultaneously (Hickendorff et al., 2018).

LPA is a model-based technique that is a version of the traditional cluster analysis. It is similar to *k*-means cluster analysis in that both methods divide individuals into categories based on response patterns (Peugh & Fan, 2013). The objective of *k*-mean cluster analysis is to quantify separation in the multivariate distance as it categorizes individuals based on response similarities that maximize between-category variance and minimizes within-category variance. LPA aims to identify the heterogeneous (*k*>1) population model that generates the data using maximum likelihood estimation (Steinley & Brusco, 2011). In *k*-means cluster analysis, an individual either is (1) or is not (0) a member of cluster *k*. In LPA, latent profile membership is estimated as a probability based on a participant's observed/indicator variable scores (Peugh & Fan, 2013).

In Figure 1, the plots, coming from hypothetical data on height, are given (Oberski, 2016). The height distribution on the left-hand side of Figure 1 is not normal; for example, when two different normal distribution groups (suppose women and men) are combined on the right-hand side, the non-normal distribution picture emerges. When modifications are made for the sub-groups in a sample, we can obtain a picture like a plot on the right. However, the distribution obtained over the total group may not show us the same distribution. Even more commonly, we may need difficult or impossible information to get at directly. LPA is, therefore, concerned with recovering such hidden (latent) groups.

Figure 1. *People's height**.



Note: Left: observed distribution. Right: women and men separate, with the total shown as a dotted line.
*Oberski, D. L. (2016). Mixture models: latent profile and latent class analysis. In J. Robertson, & M. Kaptein (Eds.), *Modern statistical methods for HCI* (pp. 275-287). Springer International.

The LPA is beneficial in examining situations where there is doubt that a defined model does not apply to all individuals. In cases where there are many variables, and there is a need to reduce them to interpretable groups, techniques such as LPA can be used to construct a meaningful relationship between variables and interpret those relations. LPA divides the observations into mutually exclusive groups when the observed variables are unrelated to each other within each class (independent), and instead of assuming that the variables come from any specific distribution, LPA allows them to follow any distribution as long as they are independent within classes (Oberski, 2016). In summary, LPA offers a parsimonious way to classify latent profiles using theoretically reasonable and particular variables (Stanley et al., 2017). LPA can also examine the relationships between class membership and external variables not used in the model (Oberski, 2016).

Variable-centered approaches that assume homogeneity in the nature of individual differences (Hickendorff et al., 2018) emphasize the relations between variables and accept that all individuals belong to the same population or come from known groups such as gender and ethnicity. On the other hand, person-centered approaches that focus on the relationships among individuals aim to group individuals into sub-groups, each containing individuals similar to each other and different from individuals in other groups (Muthén & Muthén, 2000). LPA (person-centered approach) and factor analysis (variable-centered approach) can be compared to understand LPA better. While the main purpose of the former is to find groups of individuals who are similar by using continuous observed variables, the aim of the latter is to find the smallest number of dimensions that can explain the relationships among a set of observed continuous variables (Muthén & Muthén, 2000). The difference is that factor analysis separates the covariances to show the relationships among variables, whereas LPA separates the covariances to show the relationships among individuals (Ferguson et al., 2020).

It can be seen that the use of LPA has increased considerably by applied researchers in the educational and social sciences in recent years (Ferguson et al., 2020). LPA is frequently used in modeling latent profiles/classes related to psychological structures (Bouckenooghe et al., 2019; Ferguson & Hull, 2019; Kim & Lee, 2021; Kökçam et al., 2022; Merz & Roesch, 2011; Wang et al., 2019; Wei et al., 2021; Yalçın et al., 2022) and in defining latent profile characteristics and examining the properties of those profiles in other fields (Bondjers et al., 2018; Grunschel et al., 2013; Lehmann et al., 2019; Saritepeci et al., 2022; Stanley et al., 2017; Wade et al., 2006; Williams et al., 2016).

This current study presents a brief introduction and application of LPA for researchers interested in exploring unobserved heterogeneity and integrating this type of analysis into their research, providing a helpful guide to LPA's model requirements and reporting practices, and focusing on the practical points in the analysis, proposes approaches supported by the current methodology research, and directs the researchers to additional resources for further investigation. The present application used LPA to determine the qualitatively different emotional awareness sub-groups of preservice teachers by using Emotional Awareness Questionnaire (EAQ; Rieffe et al., 2008) data. Also, covariates were integrated into the model to explore the relationships and differences between profiles (Nylund-Gibson & Masyn, 2016) by taking the Trait Emotional Intelligence Questionnaire (TEIQ; Petrides & Furnham, 2000) scores.

Emotional competence has become prominent in educational sciences, psychology, and other fields (Ashkanasy & Dasborough, 2013; Ulloa et al., 2016) as emotional competence is a fundamental part of people's social development and identifies their ability to interact and create relationships with others (Ulloa et al., 2016). Substantial evidence shows that the way of teachers' interaction with children affects their social and emotional attitudes, while emotional competencies of teachers, like emotional awareness, play a valuable role in developing positive relationships with children and contribute to forming a healthy climate in classrooms (Gottman & Declaire, 1997; Harvey & Evans, 2003; McCarthy, 2021). Therefore, teachers' emotional competencies should be supported to meet children's emotional needs. The method presented here can be used to understand teachers' emotional awareness profiles, enrich our inferences, and enhance teachers' emotional competencies.

This study, therefore, aims to demonstrate the LPA process using emotional awareness data to identify unobserved heterogeneity in a sample, identify whether emotional awareness profiles exist among preservice teachers, and evaluate predictors of profile membership. To this end, the research questions are as follows:

1) How many latent profiles exist in the EAQ data?
2) Do TEIQ scores (as covariates) predict latent profile membership?

2. METHOD

2.1. Study Group

The data came from a larger prospective research project and were used here only for illustrative purposes to demonstrate LPA, as opposed to the theoretical implications of the results. The data were collected from 184 volunteer preservice teachers in the fall and spring terms of the 2020-2021 academic year. The study group comprised 76% female and 14% male preservice teachers, and their mean age was 21.

The required sample size in LPA depends on the number of profiles and the distance between the profiles, but these are generally unknown and can only be estimated based on prior research

(Tein et al., 2013). However, there is currently no simple formula or calculator to estimate the required sample. Wurpts and Geiser (2014) suggested that sample sizes are well into the hundreds, and samples below *n*=70 are not suitable under virtually any circumstances. In this study, it is assumed that the sample size is feasible for LPA.

2.2. Data Collection Tools

Emotional Awareness Questionnaire (EAQ; Rieffe et al., 2008) aims to identify how people feel and think about their feelings. The present EAQ (30 items) was designed with a six-factor structure describing six aspects of emotional functioning; namely, (1) differentiating emotions, (2) verbal sharing of emotions, (3) not hiding emotions (formerly acting out), (4) bodily awareness of emotions, (5) attending to others' emotions, and (6) analyses of emotions. The respondents were asked to rate the degree to which each item was proper for them on a 5-point scale (*from 1 = not true to 5 = true*).

Scale items were translated into Turkish by the researchers, and Exploratory Factor Analysis (EFA) was conducted to examine the factor structure of the adapted version. According to the results of EFA, a Kaiser-Meyer-Olkin (KMO) value was found to be 0.81. Chi-square (χ^2) statistic and the result of Bartlett's test were statistically significant (χ^2 (435) = 2372.97, $p <$ .05). The data were found to have a six-factor structure with eigenvalues between 1.01 and 5.87, and the total variance explained by the factors was 49.89%. Cronbach α reliability coefficient was calculated for each sub-factor and found as 0.82, 0.71, 0.74, 0.82, 0.82, and 0.81.

Trait Emotional Intelligence Questionnaire (TEIQ; Petrides & Furnham, 2000) scores were added as covariates to the LPA model. Turkish version of TEIQ (Deniz et al., 2013) was used to measure the level of self-perception of an individual's emotional competencies. Emotional intelligence can be assessed under such four sub-factors in TEIQ as 1) emotionality, 2) well-being, 3) social, and 4) self-control. Each factor is measured by four items. Items can be responded to on a 7-point scale, ranging from "applies to me very well" (7 points) to "does not apply to me at all" (1 point). Values for each factor vary between 4 and 28. In this study, the Cronbach α reliability coefficient for each sub-factor was calculated as 0.50, 0.73, 0.64, and 0.48.

2.3. Data Analysis

2.3.1. *Latent profile analysis*

LPA (Lanza et al., 2003) is a technique for discovering latent groups in data by acquiring the probability of individuals regarding different groups. LPA thoroughly investigates the distributions of groups in the sample and determines whether those distributions are substantial or not. It might be helpful to consider if these groups are profiles of individuals as observed latent mixture components or not (Ferguson & Hull, 2019).

In LPA, the researcher works through an iterative modeling process to define the number of profiles and fits a covariate model to explore the effect of these profiles on other variables or to estimate profile membership (Bauer & Curran, 2004; Sterba, 2013). The object of LPA is to discover latent profiles (k) of individuals (i) who share a meaningful and interpretable pattern of responses on the measures of interest (j) (Marsh et al., 2009; Masyn, 2013). The joint and marginal probabilities in within-class and between-class models are used to estimate latent profiles. Within-class model is defined by two equations (Ferguson et al., 2020) as follows:

$$y_{ij} = \mu_j^{(k)} + \varepsilon_{ij} \tag{1}$$

$$\varepsilon_{ij} \sim N(0, \sigma_j^{2\,(k)}) \tag{2}$$

where $\mu_j^{(k)}$ is the model denoted mean and $\sigma_j^{2\ (k)}$ is the model denoted variance, which will vary across $j = 1 \ldots J$ outcomes and $k = 1 \ldots K$ classes or profiles, and ε_{ij} denoted the error term. The general assumption of LPA implies that outcome variables are normally distributed and locally independent* within each class (Sterba, 2013). The between-class model represents the probability of membership in a given class k:

$$p(c_i = k) = \exp\ (\omega^{(k)})\ /\ \sum_{k=1}^{K} \exp(\omega(k)) \tag{3}$$

where $\omega^{(k)}$ is a multinomial intercept and c_i is the latent classification variable for the individual. The within-class and between-class models can therefore be combined into a single model using total probability resulting in

$$f(y_i) = \sum_{k=1}^{K} p(ci\ =\ k) f(y_i\,|c_i{=}k) \tag{4}$$

which is the marginal probability density function for an individual (i) after summing across the joint within-class density probabilities for the J outcome variables, weighted by the probability of class or profile membership from equation (3). Finally, LPA results in a posterior probability for each individual are defined as

$$t_{ik} = p(c_i = k|\mathbf{y}_i) = \frac{p(ci = k) f(yi\ |ci{=}k)}{f yi} \tag{5}$$

representing the probability of an individual (i) being assigned membership (c_i) in a specific class or profile (k) given their scores on the outcome variables in the y_{i} vector. A posterior probability (t) is calculated for each individual in each profile, with values closer to 1.0 indicating a higher probability of membership in a specific profile. The more distinctions between an individual's posterior probabilities, the more certainty there is around their membership assignment (Sterba, 2013).

In general, as the number of indicators and/or latent profiles/classes increases, the number of parameters to be estimated increases; especially the number of free parameters associated with variances and covariances increases. For more parsimonious models, researchers assume that the class-specific covariance matrix is diagonal (i.e., all within-cluster covariances are equal to zero), which forces a constraint of homogeneity of variances across latent profiles. The result of these constraints is that all the latent profiles have the same form of distribution, differing only in their means (Tein et al., 2013).

2.3.2. *Steps of LPA*

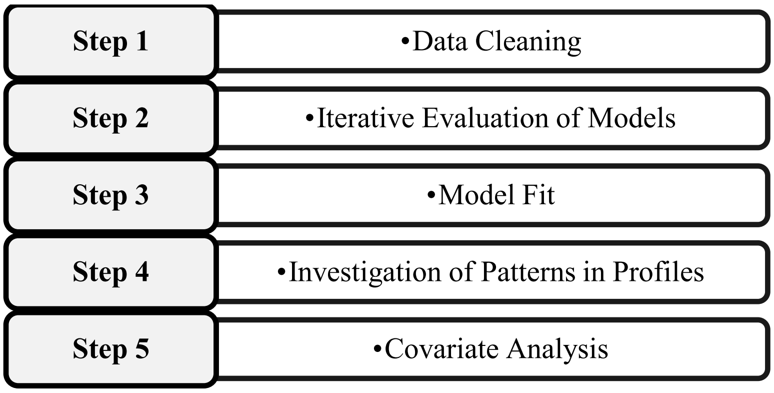
The analysis has five common steps (shown in Figure 2) as defined by Ferguson et al. (2020). *Step 1* involves data cleaning for analysis and checking for standard statistical assumptions. In the present application of LPA, the data did not contain missing values because those participants who had a missing value on one of the scales were removed from the data. However, if the data has missing values, it can be handled by full-information maximum likelihood (FIML) or multiple imputations, depending on what is best for the data (Ferguson & Hull, 2019).

Step 2 involves assessing a series of hypothetically plausible iterative LPA models, starting with one profile, and ending with the best fit of the model to the data (Hickendorff et al., 2018). Model 1 was estimated with only one profile, Model 2 with two profiles, Model 3 with three profiles, and Model 4 with four profiles to determine the best-fitting model for the data. LPA

* Local independence is a default assumption in many latent variable models but can be relaxed (Bauer, 2022). M*plus* program, by default, also imposes local independence and homogeneity across classes.

was conducted using M*plus* (8.3 version) (Muthén & Muthén, 1998-2017) with maximum likelihood estimation with robust standard errors (MLR).

Figure 2. *Five steps of Latent Profile Analysis*.*



* Ferguson, S. L., Moore, E. W., & Hull, D. M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development, 44*(5), 458-468. https://doi.org/10.1177/0165025419881721

Step 3 involves assessing models to define model fit and interpretability. One of the essential works in LPA is accurately describing the number of underlying latent profiles and correctly placing individuals into their profiles with high precision. Appropriately selecting the correct number of latent profiles is crucial because the number of profiles chosen can have a powerful impact on substantive interpretations of the results (Tein et al., 2013). Selecting the number of profiles typically involves estimating models with incremental numbers of latent profiles (e.g., 2, 3, and 4 latent classes) and choosing the number of profiles based on which model best fits the observed data. The model selection process is probably the most prominent and challenging issue. Most common methods for deciding the number of profiles fall into three categories: information criterion methods, likelihood ratio statistical test methods, and the entropy index (Nylund et al., 2007; Tein et al., 2013).

The first category, information-theoretic methods, involves Akaike's Information Criterion (AIC; Akaike, 1987) and Bayesian Information Criterion (BIC; Schwarz, 1978), which are the most commonly used indices. The AIC and BIC are based on the maximum likelihood estimates of the model parameters for deciding the most parsimonious and correct model. AIC and BIC are used for model selection, with lower values representing the retained model (Masyn, 2013).

The second category involves likelihood ratio statistic tests (LRTs) that compare the relative fit of two models that differ by a set of parameter restrictions (Tein et al., 2013). To illustrate, Lo, Mendell, and Rubin (LMR-LRT) is used to compare models in a similar context to the χ^2 difference test in other model testing analyses (Lo et al., 2001); LMR-LRT evaluates significance across differences in degrees of freedom and helps determine when additional profiles are not improving the fit or discrimination of the model. Thus, a nonsignificant LMR-LRT suggests that the more parsimonious model fits better (Ferguson & Hull, 2019). The bootstrap likelihood ratio test (BLRT) and Vuong-Lo-Mendell-Rubin (VLMR-LRT) can be used to compare the fit of one model (k) compared to a model with one less class (k-1). BLRT uses parameter estimation methods to create multiple bootstrap samples representing the sampling distribution (Masyn, 2013). A statistically significant BLRT indicates that the current model fits better than a model with a k-1 class. For LMR-LRT, VLMR-LRT, and BLRT, a small probability value (e.g., $p < .05$) indicates that the k-class model provides a significantly better fit to the observed data than the $k-1$ class model does (Whittaker & Miller, 2021).

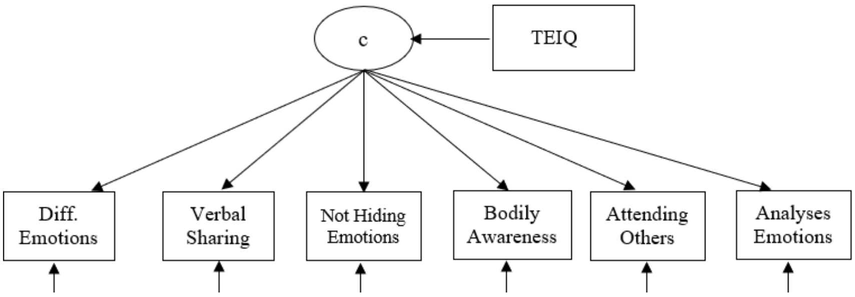
The third category is the measure of entropy. The entropy index is based on the uncertainty of classification (Celeux & Soromenho, 1996), while the entropy index scales to the interval (0, 1). A higher value of entropy represents a better fit; values > 0.80 indicate that the latent profiles are highly discriminating (Tein et al., 2013).

In addition to model fit indices, evaluating the reasonableness of an LPA model is necessary to provide the final model and underlying profiles that represent interpretable and meaningful groupings of individuals. Profiles containing less than 5% of the sample may be spurious, and the relevance of such profiles should be carefully considered and examined for interpretability (Marsh et al., 2009).

Step 4 involves interpreting the retained model by examining patterns of the profiles and weights of variables included in each profile. The means and standard deviations of variables used to create the profiles are conditioned and presented for each profile. It may help report LPA to provide names for the profiles based on the observed differences in indicator variables. Correct naming of profiles provides accuracy and clarity in generalizing and interpreting results (Ferguson & Hull, 2019).

Step 5 involves conducting a covariate analysis. This step should be included when (a) LPA analysis indicates that there are profiles worth interpreting further and (b) there is a theoretical reason to evaluate the impact of the covariates on the profiles (Ferguson & Hull, 2019). Examining relationships with covariates provides additional information on the latent profiles and how the covariate variables may have differing effects on these profiles. A three-step approach (Vermunt, 2010) is used for the inclusion of covariates in the LPA. The first step of the three-step approach is determining the number of latent profiles without including the covariates in the model (Marsh et al., 2009); in the second step, the individuals' class probabilities are used to specify their membership probability into each latent profile; and in the third step, the logit values for the most likely class are regressed on covariate variables, considering the misclassification in the second step (Asparouhov & Muthen, 2014). Using the three-step approach means that indicators for the profiles are present in the model with the covariates during data analysis, as shown in Figure 3.

Figure 3. *LPA model with covariate.*



Note. TEIQ=Emotional intelligence scores; Diff. Emotions=Differentiating Emotions.
Differentiating emotions, verbal sharing, not hiding emotions, bodily awareness, attending to others, and analyses of emotions are observed/indicators of emotional awareness.

Figure 3 shows the observed/indicator and covariate variables for the EAQ construct. TEIQ scores were added by regressing the latent profile membership into the model as a covariate of latent class *c* in Figure 3. In this study, first, basic LPA models were tested and examined to identify the presence of latent profiles of emotional awareness (research question 1), and then, LPA models with covariates were tested and examined to evaluate the effects of covariates for defining latent profiles (research question 2).

3. RESULTS

LPA results are given in accordance with the steps followed, and the research questions asked. In Step 1, the data were cleaned, and participants were removed from the analysis if values on all variables in the study were missing. Therefore, the results involve LPA steps from two to five.

3.1. Results of Research Question-1

Research question 1 involves results from the second to the fourth step of LPA. In Step 2, a series of LPA models were evaluated, starting with one profile (Model 1) and ending with a model with four profiles (Model 4).

Step 3 involves evaluating model fit to identify latent profiles. Model 3 was retained as the best-fitting model to the data based on the lower AIC and BIC values, high entropy, and the significant LMR-LRT, while the smallest class contained more than 5% of the sample (Table 2). BIC was marginally lower for Model 3 compared to Model 4. The entropy for Model 3 was 0.77. The LMR-LRT, VLMR-LRT, and BLRT tests were significant for Model 3, which means the three-profile model is better than the two-profile model. These results showed that adding new classes to the model, from the one-class to the three-class model, improves the model-data fit. However, adding a class to the three-profile model did not improve the model-data fit because LMR-LRT was not significant for Model 4, which means the more parsimonious Model 3 had a better fit than that of the less parsimonious model. The smallest profile in Model 3 comprised 9% (*n*=17) of the sample. It was therefore concluded that the three-profile model better fits the data under the interpretability and parsimonious principle.

Table 2. *Model fit summary of LPA models.*

Model Fit Statistics	Model 1	Model 2	Model 3*	Model 4
AIC	6047.74	5945.87	5907.27	5898.07
BIC	6086.32	6006.95	5990.86	6004.16
Entropy	*	0.67	0.77	0.70
Smallest class %	*	49	9	8
LMR-LRT p-value	*	0.00	0.02	0.84
VLMR-LRT p-value	*	0.00	0.02	0.83
BLRT p-value	*	0.00	0.00	0.06

Note. *n*=184; *p*-value < .05 *Retained model for the emotional awareness data

In Step 4, the retained model was interpreted by examining the patterns of the latent profiles. As the results of the three-profile model are given in Table 3, it can be seen that the standardized means used to create the classes were presented for each profile, and all were found to be statistically significant. *Profile 1* contained preservice teachers with the lowest level of differentiating emotions, verbal sharing, not hiding emotions, and bodily awareness (lower values indicate that more bodily symptoms accompany emotions), which was referred to as "Introverted". *Profile 2* contained preservice teachers with the mid-level of differentiating emotions, verbal sharing, not hiding emotions, the highest level of bodily awareness, the lowest level of attending to others, and analyses emotions, so it was referred to as "Less Sensitive to Others' Emotions". *Profile 3* contained preservice teachers with the highest level of differentiating emotions, verbal sharing, not hiding emotions, the mid-level of bodily awareness, the highest level of attending to others, and analyses of emotions, so it was referred to as "Extroverted".

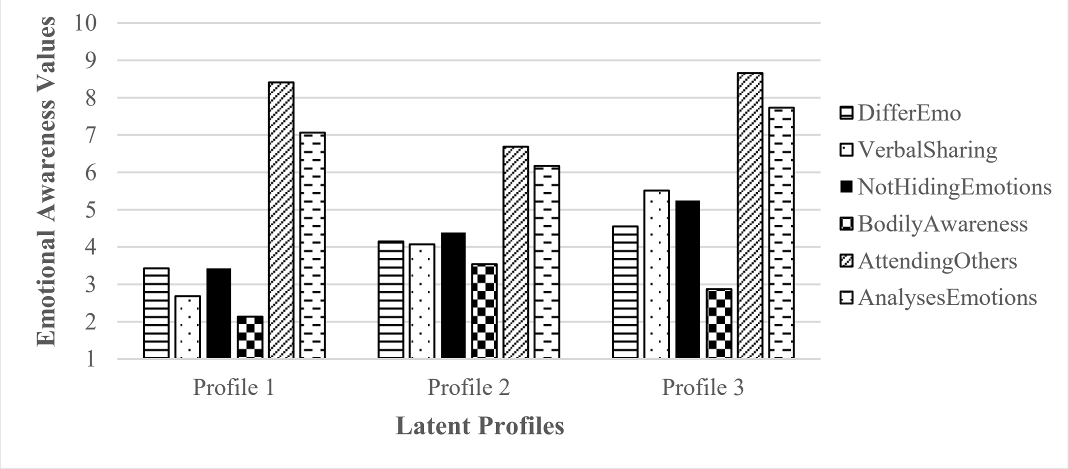
Additionally, the classification uncertainty value entropy was calculated as 0.77. This result shows that the retained three-profile model was effective in assigning individuals to the correct latent profiles. The latent profile membership of each participant was calculated based on the posterior class probabilities, which represent the probability of being in each of the *k* latent classes based on observed responses to the items. It was seen that classification probabilities in the three-profile model were 0.80 or greater (0.80, 0.88, 0.95), indicating that participants were assigned to corresponding latent profiles with high probabilities. This result supported the usefulness of the three-profile model in assigning individuals to the correct classes.

Table 3. *Mean values of observed variables for three-profiles model.*

Indicator Variables	Profile 1	Profile 2	Profile 3
Differentiating Emotions	3.43	4.14	4.55
Verbal Sharing	2.68	4.07	5.50
Not Hiding Emotions	3.42	4.38	5.24
Bodily Awareness	2.12	3.54	2.87
Attending to Others	8.40	6.68	8.65
Analyses Emotions	7.05	6.17	7.73
Class Proportions	9% (*n* = 17)	35% (*n* = 65)	56% (*n* = 102)

Figure 4 involved plots for comparing profiles on indicator variables. While *Profiles 1* and *3* contained preservice teachers with high scores in attending to others and analyses emotions, *Profile* 2 contained preservice teachers with lower scores in attending to others, analyses emotions, and higher scores in bodily awareness compared to those of other profiles.

Figure 4. *Histograms for latent profiles.*



Note. Profile 1=Introverted, Profile 2= Less sensitive to others' emotions, Profile 3= Extroverted DifferEmo: differentiating emotions, VerbalSharing: verbal sharing of emotions, NotHidingEmotions: not hiding emotions (formerly acting out), BodilyAwareness: bodily awareness of emotions, AttendingOthers: attending to others' emotions, and AnalysesEmotions: analyses of emotions.

3.2. Results of Research Question-2

Research question 2 involves the results of covariate analysis (Step 5). TEIQ scores were added to the model by regressing the latent profile. *Profile 3* (Extroverted) was used as the reference group for model comparisons.

Odds ratios were computed to evaluate differences in the likelihood of profile membership based on covariate scores. TEIQ sub-factors were emotionality, well-being, sociability, and self-control. Odds ratios demonstrating the likelihood of profile membership based on covariate compared to *Profile 3* (Extroverted) are presented in Table 4. Some of the TEIQ sub-factors produced significant differences across profiles ($p < .05$). Positive coefficients indicated that the probability of participants with high related TEIQ sub-factor scores tends to be other profiles as compared to the reference profile (*Profile 3 - Extroverted)*. Regarding the significant and negative coefficients, it could be implied that participants with high levels of emotionality and well-being were less likely to be in *Profile 1 (Introverted)* and *Profile 2 (Less sensitive to others' emotions)* compared to the reference profile. Besides, participants with high sociability values were less likely to be in *Profile 2* relative to *Profile 3;* however, the self-control sub-factor had no significant effect.

Table 4. *Covariate analysis results for the three-profile model.*

Covariate Variables	Latent Profiles	
	Profile 1 (slope/odds ratio) Introverted	Profile 2 (slope/odds ratio) Less sensitive to others' emotions
Emotionality	-0.16 / 0.85*	-0.25 / 0.78*
Well-being	-0.21 / 0.81*	-0.15 / 0.86*
Sociability	-0.15 / 0.86	-0.15 / 0.86*
Self-control	0.04 / 1.04	0.03 / 1.03

Note. *p < .05; Reference class= Profile 3 (Extroverted)

4. DISCUSSION and CONCLUSION

In this paper, we provide an overview of LPA and highlight the strengths of this analytic approach, which is a member of latent variable mixture models and uses continuous data collected from cross-sectional measurement points (Berlin et al., 2014). A step-by-step LPA guide was provided illustrating the methodology which was used to determine the number of meaningful latent classes and their patterns to advance our understanding of preservice teachers' capabilities in relation to emotional awareness.

Collected as a part of a larger research project focusing on social and emotional competencies of preservice teachers, emotional awareness data were used 1) to predict the latent profile construct underlying the data and to test if some of the profile differences could be explained by the emotionality, well-being, sociability, and self-control as covariates, and 2) to interpret the resulting emotional awareness profiles as they pertained to the desired qualifications in the teaching profession. The results showed that, based on their EAQ scores, preservice teachers could be classified into three distinct profiles, namely Introverted (9%), Less Sensitive to Others' Emotions (35%), and Extroverted (56%), suggesting that there were sub-groups of preservice teachers having distinct characteristics and needs. According to the results, only up to half of the teachers in the sample were identified as having the professionally desired emotional awareness levels. Furthermore, some of the TEIQ sub-factors that were tested as covariates were found to play an important role in profile memberships. For example, it was found that preservice teachers with higher well-being and emotionality self-ratings were less likely to be introverted and less sensitive to others' profiles compared to extroverted profile. Overall, our findings indicate that we need to consider the added value of utilizing theoretically meaningful hypotheses and covariate variables in order to investigate the profile patterns of teachers in a detailed way.

The present study also highlights the importance of recovering hidden sub-groups within the sample of preservice teachers. LPA can be beneficial, especially for gaining a better understanding of the characteristics of the target populations. Teachers with higher social and emotional capabilities are expected to show more awareness about their own emotions, discriminate between their feelings and those of others, monitor and regulate their internal processes, and understand more accurately the causes of emotions in themselves and the children they work with in comparison to those with little social or non-emotionally capabilities (Jennings & Greenberg, 2009). Because of these capabilities, emotionally aware teachers are expected to implement positive strategies and cultivate self-awareness skills to understand and reflect on the emotional difficulties that underlie children's behavior (Ulloa et al., 2016) since developing emotional awareness competencies has been reported to reduce inappropriate behaviors in the classroom, reduce stress, and improve achievement (McCarthy, 2021).

Although only several teacher certification programs to date are known to emphasize social-emotional competencies in their list of priority competencies (McCarthy, 2021), many studies recognize the importance of integrating social-emotional skills into teacher education programs (Ulloa et al., 2016). Our results confirm that preservice teachers differ qualitatively concerning their emotional awareness capabilities, and also enhancing teacher training programs to diagnose their social and emotional capabilities can set the basis for designing or modifying coursework and other activities serving the needs of those in different stages of their social-emotional development. Our results, therefore, indicate that some preservice teachers appear to be in the less-than-ideal emotional awareness profiles and could use the help of additional training programs and other aids.

Some limitations need to be noted regarding the present study. This study is limited to university students, which may have affected the generalizability of the results. In addition, emotional intelligence sub-factor scores were considered for the classification of emotional awareness profiles, and due to this, understanding of memberships of emotional awareness groups may have remained limited. Since this research was exploratory, it is necessary to examine its validity through confirmatory analyses in future studies. Evaluation of item and scale parameter estimates can also inform other researchers, especially when making inferences about the potential use of alternative model covariates, for there could be different effects across different latent classes (Whittaker & Miller, 2021). For instance, the effects of covariates on class membership might improve model performance in terms of a correct class assignment (Lubke & Muthen, 2005). Although greatly useful as a methodology, researchers are recommended to formulate LPA models using theoretically meaningful latent class constructs and covariates to the extent possible.

This paper uses LPA to diagnose and describe preservice teachers' latent profile differences concerning their emotional awareness levels and social-emotional skills in general. Also, it illustrates that adding predictor variables as covariates to the LPA models may help discover relationships and other inherent differences between latent groups (Bouckenooghe et al., 2019; Hill et al., 2006; Nylund-Gibson & Masyn, 2016; Stanley et al., 2017). By utilizing LPA, detailed information can be obtained about qualitative individual differences related to the particular construct of interests to further understand preservice teachers' characteristics and determine their strengths and weaknesses. Thus, through timely diagnostics and proper curricula or program improvement targeting specific needs and skills, existing teacher training programs can be updated to empower future teachers.

Acknowledgments

This work was partially supported by the Gazi University Scientific Research Council [grant number 04/2019-01] and by Grant [120K122] from the Scientific and Technological Research Council of Turkey (TUBITAK).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The studies involving human participants are reviewed and approved by the Gazi University Ethics Committee (Ethics Committee Number: 77082166-604.01.02).

Authorship Contribution Statement

Esra Sozer-Boz: Investigation, Resources, Data Collection, Visualization, Software, Methodology, and Writing-original draft. **Derya Akbas**: Investigation, Data Collection, Methodology, Software, and Validation. **Nilufer Kahraman**: Supervision and Validation.

Orcid

Esra Sozer-Boz https://orcid.org/0000-0002-4672-5264
Derya Akbas https://orcid.org/ 0000-0001-9852-4782
Nilufer Kahraman https://orcid.org/ 0000-0003-2523-0155

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317-332. https://doi.org/10.1007/BF02294359

Ashkanasy, N.M. & Dasborough, M.T. (2003) Emotional awareness and emotional intelligence in leadership teaching. *Journal of Education for Business, 79*(1), 18-22. https://doi.org/10.1080/08832320309599082

Asparouhov, T., & Muthen, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 329-341. https://doi.org/10.1080/10705511.2014.915181

Bauer, J. (2022). A primer to Latent Profile and Latent Class Analysis. In M. Goller, E. Kyndt, S. Paloniemi & C. Damşa (Eds.), *Methods for researching professional learning and development: Challenges, applications, and empirical illustrations* (pp. 243-268). Springer Cham. https://doi.org/10.1007/978-3-031-08518-5

Bauer, D.J., & Curran, P.J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*(1), 3-29. https://doi.org/10.1037/1082-989X.9.1.3

Berlin, K.S., Parra, G.R., & Williams, N.A. (2014). An introduction to latent variable mixture modeling (part 2): longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology, 39*(2), 188-203. https://doi.org/10.1093/jpepsy/jst085

Bondjers, K., Willebrand, M., & Arnberg, F.K. (2018). Similarity in symptom patterns of posttraumatic stress among disaster-survivors: a three-step latent profile analysis. *European Journal of Psychotraumatology, 9*(1). https://doi.org/10.1080/20008198.2018.1546083

Bouckenooghe, D., Clercq, D.D., & Raja, U. (2019). A person-centered, latent profile analysis of psychological capital. *Australian Journal of Management, 44*(1), 91-108. https://doi.org/10.1177/0312896218775153

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(2), 195-212. https://doi.org/10.1007/BF01246098

Deniz, M.E., Özer, E., & Işık, E. (2013). Trait Emotional Intelligence Questionnaire–Short Form: Validity and reliability studies. *Education and Science, 38*(169), 407-419.

Ferguson, S.L., & Hull, D.M. (2019). Exploring science career interest: Latent profile analysis of high school occupational preferences for science. *Journal of Career Development, 46*(5), 583-598. https://doi.org/10.1177/0894845318783873

Ferguson, S.L., Moore, E.W., & Hull, D.M. (2020). Finding latent groups in observed data: A primer on latent profile analysis in Mplus for applied researchers. *International Journal of Behavioral Development, 44*(5), 458-468. https://doi.org/10.1177/0165025419881721

Gibson, W.A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24*, 229-252. https://doi.org/10.1007/BF02289845

Gottman, J. & Declaire, J. (1997). *Raising an emotionally intelligent child: The heart of parenting.* Simon & Schuster.

Grunschel, C., Patrzek, J., & Fries, S. (2013). Exploring different types of academic delayers: A latent profile analysis. *Learning and Individual Differences, 23*, 225-233. https://doi.org/10.1016/j.lindif.2012.09.014

Harvey, S. & Evans, I.M. (2003). Understanding the emotional environment of the classroom. In D. Fraser & R. Openshaw (Eds.), *Informing our practice* (pp. 182-195). Kanuka Grove.

Hickendorff, M., Edelsbrunner, P.A., Schneider, M., Trezise, K., & & McMullen, J. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences, 66,* 4-15. https://doi.org/10.1016/j.lindif.2017.11.001

Hill, A.L., Degnan, K.A., Calkins, S.D., & Keane, S.P. (2006). Profiles of externalizing behavior problems for boys and girls across preschool: The roles of emotion regulation and inattention. *Developmental Psychology, 42*(5), 913-928. https://doi.org/10.1037/0012-1649.42.5.913

Jennings, P.A., & Greenberg, M.T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research, 79*(1), 491-525. http://dx.doi.org/10.3102/0034654308325693

Jung, T., & Wickrama, K.A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*(1), 302-317. https://doi.org/10.1111/j.1751-9004.2007.00054.x

Kim, S., & Lee, Y. (2021). Examining the profiles of school violence and their association with individual and relational covariates among South Korean children. *Child Abuse & Neglect, 118*. https://doi.org/10.1016/j.chiabu.2021.105155

Kökçam, B., Arslan, C., & Traş, Z. (2022). Do psychological resilience and emotional intelligence vary among stress profiles in university students? A latent profile analysis. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.788506

Lanza, S.T., & Rhoades, B.L. (2013). Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science, 14*, 157-168. https://doi.org/10.1007/s11121-011-0201-1

Lanza, S.T., Flaherty, B.P., & Collins, L.M. (2003). Latent class and latent transition analysis. In J.A. Schinka, & W.A. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (pp. 663-685). Wiley.

Lehmann, R.J., Neumann, C.S., Hare, R.D., Biedermann, J., Dahle, K.P., & Mokros, A. (2019). A latent profile analysis of violent offenders based on PCL-R factor scores: Criminogenic needs and recidivism risk. *Frontiers in Psychiatry, 10,* 627. https://doi.org/10.3389/fpsyt.2019.00627

Lo, Y., Mendell, N.R., & Rubin, D.B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*(3), 767-778. https://doi.org/10.1093/biomet/88.3.767

Lubke, G.H., & Muthen, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21-39. https://doi.org/10.1037/1082-989X.10.1.21

Marsh, H.W., Ludtke, O., Trautwein, U., & Morin, A.J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling, 16*, 191-225. https://doi.org/0.1080/10705510902751010

Masyn, K.E. (2013). Latent class analysis and finite mixture modeling. In T.L. (Eds.), *The Oxford handbook of quantitative methods* (pp. 551-611). Oxford University.

McCarthy, D. (2021). Adding social emotional awareness to teacher education field experiences. *The Teacher Educator*. https://doi.org/10.1080/08878730.2021.1890291

Merz, E.L., & Roesch, S.C. (2011). A latent profile analysis of the Five Factor Model of personality: Modeling trait interactions. *Personality and Individual Differences, 51*(8), 915-919. https://doi.org/10.1016/j.paid.2011.07.022

Muthén, B. (2007). Latent variable hybrids: Overview of old and new methods. In G.R. Hancock & K.M. Samuelsen (Eds.), *Advances in latent variable mixture modeling* (pp. 1-24). Information Age.

Muthén, B., & Muthén, L.K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882-891. https://doi.org/10.1111/j.15300277.2000.tb02070.x

Muthén, L.K., & Muthén, B.O. (1998-2017). *Mplus user's guide (8th Edition).* Muthén & Muthén.

Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569. https://doi.org/10.1080/10705510701575396

Nylund-Gibson, K., & Masyn, K.E. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling, 23*, 782-797. https://doi.org/10.1080/10705511.2016.1221313

Oberski, D.L. (2016). Mixture models: latent profile and latent class analysis. In J. Robertson, & M. Kaptein (Eds.), *Modern statistical methods for HCI* (pp. 275-287). Springer.

Petrides, K.V., & Furnham, A. (2000). Gender differences in measured and self-estimated trait emotional intelligence. *Sex Roles, 42*(5), 449-461. https://doi.org/10.1023/A:1007006523133

Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: A Monte Carlo simulation. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(4), 616-639. https://doi.org/10.1080/10705511.2013.824780

Rieffe, C., Oosterveld, P., Miers, A.C., Terwogt, M.M., & & Ly, V. (2008). Emotion awareness and internalising symptoms in children and adolescents: The Emotion Awareness Questionnaire revised. *Personality and Individual Differences, 45*(8), 756-761. https://doi.org/10.1016/j.paid.2008.08.001

Roesch, S.C., Villodas, M., & Villodas, F. (2010). Latent class/profile analysis in maltreatment research: A commentary on Nooner et al., Pears et al., and looking beyond. *Child Abuse & Neglect, 34*(3), 155-160. https://doi.org/10.1016/j.chiabu.2010.01.003

Saritepeci, M., Yildiz-Durak, H., & Atman-Uslu, N. (2022). A Latent Profile Analysis for the Study of Multiple Screen Addiction, Mobile Social Gaming Addiction, General Mattering, and Family Sense of Belonging in University Students. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-022-00816-y

Schwarz, G. (1978). Estimating the dimension of a model. *The Annuals of Statistics, 6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Snow, R.E. (1986). Individual differences and the design of educational programs. *American Psychologist, 41*, 1029-1039. https://doi.org/10.1037/0003-066X.41.10

Stanley, L., Kellermans, F.W., & Zellweger, T.M. (2017). Latent profile analysis: Understanding family firm profiles. *Family Business Review, 30*(1), 84-102. https://doi.org/10.1177/0894486516677426

Steinley, D., & Brusco, M.J. (2011). Evaluating mixture modeling for clustering: Recommendations and cautions. *Psychological Methods, 16*, 63-79. https://doi.org/10.1037/a0022673

Sterba, S.K. (2013). Understanding linkages among mixture models. *Multivariate Behavioral Research, 48*, 775-815. https://doi.org/10.1080/00273171.2013.827564

Tein, J.Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling, 20*, 640-657. https://doi.org/10.1080/10705511.2013.824781

Ulloa, M., Evans, I., & Jones, L. (2016). The effects of emotional awareness training on teachers' ability to manage the emotions of preschool children: An experimental study. *Escritos de Psicología, 9*(1), 1-14. https://doi.org 10.5231/psy.writ.2015.1711

Vermunt, J.K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

Wade, T.D., Crosby, R.D., & Martin, N.G. (2006). Use of latent profile analysis to identify eating disorder phenotypes in an adult Australian twin cohort. *Arch Gen Psychiatry, 63*(12),1377–1384. https://doi.org/10.1001/archpsyc.63.12.1377

Wang, Y., Su, Q., & Wen, Z. (2019). Exploring latent profiles of empathy among chinese preschool teachers: A person-centered approach. *Journal of Psychoeducational Assessment, 37*(6), 706-717. https://doi.org/10.1177/0734282918786653

Wei, M., Mallinckrodt, B., Arterberry, B.J., Liu, S., & Wang, K.T. (2021). Latent profile analysis of interpersonal problems: attachment, basic psychological need frustration, and psychological outcomes. *Journal of Counseling Psychology, 68*(4), 467-488. https://doi.org/10.1037/cou0000551

Whittaker, T.A., & Miller, J.E. (2021). Exploring the enumeration accuracy of cross-validation indices in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(3), 376-390. https://doi.org/10.1080/10705511.2020.1802280

Williams, K.E., Nicholson, J.M., Walker, S., & Berthelsen, D. (2016). Early childhood profiles of sleep problems and self-regulation predict later school adjustment. *British Journal of Educational Psychology, 86*(2), 331-350. https://doi.org/10.1111/bjep.12109

Wurpts, I.C., & Geiser, C. (2014). Is adding more indicators to a latent class analysis beneficial or detrimental? Results of a Monte-Carlo study. Frontiers in Psychology, *5*, 920. http://dx.doi.org/10.3389/psyg.2014.00920

Yalçın, İ., Can, N., Mançe Çalışır, Ö., Yalçın, S., & Çolak, B. (2022). Latent profile analysis of COVID-19 fear, depression, anxiety, stress, mindfulness, and resilience. Current Psychology, *41*, 459-469. https://doi.org/10.1007/s12144-021-01667-x

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 145–159

https://doi.org/10.21449/ijate.1135368

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

A Comparison of the efficacies of differential item functioning detection methods

Munevver Basman[1,*]

[1]Marmara University, Faculty of Education, Department of Educational Sciences, Türkiye

ARTICLE HISTORY

Received: June 24, 2022

Accepted: Mar. 02, 2023

Keywords:

Crossing simultaneous item bias test,
Differential item functioning,
Logistic Regression,
Lord's chi-square,
Mantel-Haenszel,
Raju's area measure.

Abstract: To ensure the validity of the tests is to check that all items have similar results across different groups of individuals. However, differential item functioning (DIF) occurs when the results of individuals with equal ability levels from different groups differ from each other on the same test item. Based on Item Response Theory and Classic Test Theory, there are some methods, with different advantages and limitations to identify items that show DIF. This study aims to compare the performances of five methods for detecting DIF. The efficacies of Mantel-Haenszel (MH), Logistic Regression (LR), Crossing simultaneous item bias test (CSIBTEST), Lord's chi-square (LORD), and Raju's area measure (RAJU) methods are examined considering conditions of the sample size, DIF ratio, and test length. In this study, to compare the detection methods, power and Type I error rates are evaluated using a simulation study with 100 replications conducted for each condition. Results show that LR and MH have the lowest Type I error and the highest power rate in detecting uniform DIF. In addition, CSIBTEST has a similar power rate to MH and LR. Under DIF conditions, sample size, DIF ratio, test length and their interactions affect Type I error and power rates.

1. INTRODUCTION

Tests are tools that contain systematic processes used to evaluate latent traits (Linn & Gronlund, 2000). With the results obtained from the tests, groups with different traits can be compared, and various decisions can be made based on the comparison results. However, if the test items are biased in favor of a group and not fair, the validity of the test is affected (Kane, 2006; Messick, 1989). For this reason, studies on the reliability and validity of the tests are carried out.

One way to ensure the validity of the tests is to check that all items work similarly across different groups of individuals. Differential item functioning (DIF) occurs, however, when individuals with equal ability levels from various groups perform differently on the same test item. In other words, DIF is the differentiation of the probability of subgroups with the same ability to correctly answer the item (Gao, 2019; Hambleton et al., 1991). While determining DIF in bias studies, two groups can be studied as the focus and the reference groups. The focus group is the one in which the negative situations of individuals with the same ability are examined while responding to the item. The group to which the focus group is compared is

*CONTACT: Munevver Başman ✉ munevver.rock@gmail.com Marmara University, Faculty of Education, Department of Educational Sciences, Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

called the reference group (Zumbo, 1999). The focus group is also called the minority, and the reference group as the majority (de Ayala, 2009). When comparing the item parameters and the item characteristic curves (ICC) of the groups, it is checked whether they are different.

DIF occurs in two forms: uniform DIF and non-uniform DIF (Mellenbergh, 1983). The item examined in the uniform DIF has a situation where a certain group works in favor of the other group at every ability level. In other words, it is a situation where the percentage of a group answering an item correctly at each ability level is consistently high (Osterlind & Everson, 2009). The ICCs of both groups are different and do not overlap with each other. Uniform DIF is indicated when item difficulty (b-parameters) differs between groups (reference and focus group). In non-uniform DIF, the item studied is in favor of one group in a certain skill level range, while it works in favor of the other group in another ability range (Camilli & Shepard, 1994; Hambleton et al., 1993; Swaminathan & Rogers, 1990). The ICC of both groups are different, but they overlap at some point on the ability (theta) scale. Non-uniform DIF is detected when item discrimination (a-parameters) or both a and b parameters differ across groups.

DIF detection methods are basically classified according to the Classical Test Theory (CTT), which takes into account the observed score group, and Item Response Theory (IRT), which takes into account the latent variable group. Since the test score in the CTT is dependent on the item sample, there are limitations in the generalization of the DIF results. Therefore, there are trends toward IRT in later studies (Embretson & Reise, 2000; Hambleton et al., 1993). When DIF determination methods according to IRT and CTT are compared, the estimation of the item parameters with IRT gives more meaningful results than the CTT, the differences in item functions can be defined more meaningfully by plotting the differences in the IRT compared to the CTT, and it is easier with the IRT than the CTT, to understand whether the item shows DIF or not (Camilli & Shepard, 1994; Narayanan & Swaminathan, 1996). However, DIF detection methods based on IRT require large sample size and assumptions may be difficult to meet in practice (Narayanan & Swaminathan, 1994). DIF determination methods based on CTT can be preferred because CTT can also be used in small samples and assumptions are easier to meet in practice than IRT.

According to CTT, analysis of variance, chi-square, transformed item difficulty, the Mantel-Haenszel (MH) method, and the Logistic Regression (LR) procedure are some methods for detecting DIF. Some DIF detection methods based on IRT are Lord's chi-square (LORD), Raju's area measure (RAJU), the IRT Likelihood Ratio test (IRT-LR), Lord's IRT Wald test, the crossing simultaneous item bias test (CSIBTEST), and the Multiple Indicators Multiple Causes (MIMIC) model (Camilli & Shepard, 1994; Gao, 2019; Oshima & Morris, 2008). This research compares MH, LR, LORD, RAJU, and CSIBTEST methods. MH and LR methods among CTT methods are the most used methods in research due to their ease of use and interpretation (Kelecioğlu et al., 2014). Among the IRT methods, LORD based on chi-square method, RAJU based on ICC and CSIBTEST not requiring item calibration were chosen because they use different procedures.

Mantel-Haenszel method, proposed by Holland and Thayer (1988), is a test statistic based on chi-square. In this method, two levels are used for the item score variable (correct and incorrect response), two levels are used for group membership (focal and reference groups), and k levels are used for the matching variable. It is tested whether the probability of having the correct response for an item at a given level of the matching variable differs between the groups across all k levels of the matching variable (Dorans & Holland, 1992). The MH statistic based on chi-square is computed and logarithmic transformation is applied to facilitate the interpretation of MH results.

Logistic Regression, proposed by Swaminathan and Rogers (1990), can detect uniform and non-uniform DIF within dichotomous data. While determining the DIF with this method, a likelihood ratio is used (Camilli, 2006). Group belonging and total test score are the independent variables, and item score (0,1) is the dependent variable. It uses the total test score to estimate the traits of reference and focal groups and compares their response probabilities considering their ability differences.

Lord's chi-square, proposed by Lord (1980), is used to simultaneously check the differences in the item parameters between focal and reference groups. The chi-square statistic is calculated using item parameter differences and the variance-covariance matrix for these differences. A decision is made whether to reject or not reject the null hypothesis of no DIF by comparing the chi-square statistic with a critical value.

Raju's area measure, proposed by Raju (1988), detects DIF considering item characteristic curves. ICC for reference and focal groups are drawn according to the correct response probability, and the areas between these curves are compared with each other.

Simultaneous item bias test (SIBTEST) uses a latent score and does not need item calibration even though it is based on the IRT framework. The crossing simultaneous item bias test (CSIBTEST), proposed by Li and Stout (1996), is an extension of SIBTEST (Shealy & Stout, 1993). It is capable of detecting both uniform and non-uniform DIF, while SIBTEST can detect only uniform DIF.

In the literature, studies exist about the performances of DIF detection methods considering some variables. Holmes Finch and French (2007) compared SIBTEST, LR, IRT-LR, and confirmatory factor analysis (CFA) changing different factors. They found no significant differences in Type I error rates within the methods across the values used for the underlying model, group ability, and sample size. In addition, they found that power rates increased with increasing sample sizes and decreased with decreasing percentages of DIF for LR and IRT-LR. Güler and Penfield (2009) compared LR and a combination of MH and Breslow-Day (BD) procedures called the combined decision rule (CDR) to simultaneously detect both uniform and nonuniform DIF under the condition of different sample sizes and unequal ability distributions for focal and reference groups. Type I error rates and CDR and LR power rates were higher when the sample size was larger. DeMars (2009), Li et al. (2012) and Erdem Keklik (2014) compared MH and LR methods under different conditions. Type I error rates of MH and LR were found to be similar when the reference and focus group ability distributions showed a unit normal distribution. Kim (2010) compared MH, LR, LORD, and the Differential Functioning Item and Test (DFIT). A larger sample size inflated all methods' Type I error rates, and a longer test inflated the Type I error rates of MH and LR. Lopez (2012) compared the efficacy of CSIBTEST, IRT-LR, and LR. LR showed the highest predictive power and the lowest average Type I error rate. IRT-LR and CSIBTEST showed higher values than the nominal alpha level of .05. Atalay Kabasakal et al. (2014) compared the Type I errors and powers of MH, SIBTEST, and IRT-LR methods by using different values for test length, sample size, percentage of DIF, ability differences between groups, and underlying models. Type I error of SIBTEST and power rates of MH had the highest values. The factors' main and interaction effects can differentiate the methods' power and Type I error rates. Gao (2019) compared MH, LR, MIMIC model, Lord's IRT-based Wald test, IRT-LR, and a Randomization Test based on an R-square change statistic. The MIMIC model had the highest power rates. The LR had higher Type I error rates for larger sample sizes and shorter tests.

When the studies were examined, it was seen that the performances of DIF determination methods were examined by considering some variables. Although many DIF detection models have been developed and extensively studied in binary data, there are still ongoing studies in the literature on the limitations and advantages of these models and under what conditions they can be used for which data. On the other hand, it is seen in the literature that comparison studies are done under limited conditions due to their nature (Jodoin & Gierl, 2001; Li et al., 2012; Narayanan & Swaminathan, 1994; Roussos & Stout, 1996). Within the scope of this study, Type I error, and power ratios of methods based on CTT and IRT used in determining DIF were tried to be determined by considering both the main effects and the interaction effects of various conditions. From this point of view, it contributes to the field since the methods used, the conditions used and their levels are differentiated, and also the interaction effects of the factors are discussed together with the main effects. This is the first study in the literature that compares MH, LR, LORD, RAJU, and CSIBTEST methods at the same time considering different sample sizes, test lengths and proportions of DIF items.

In examining the performance of DIF methods, it was necessary to examine the uniform DIF determination processes, which commonly occur in real situations, under the conditions of ability distribution, sample size and sample size ratios, which are especially used by comparison criteria such as Type I error and which can affect the results of DIF analysis. For these reasons, in the presence of a uniform DIF underlying the 3PL model, this paper answers the following questions:

a) How do the Type I error rates of MH, LR, LORD, RAJU, and CSIBTEST methods change in conditions where the sample size is 500 and 2000; test length is 10, 20 and 30; percentage of items showing DIF is 10% and 20%?

b) How do the statistical power levels of MH, LR, LORD, RAJU, and CSIBTEST methods change in conditions where the sample size is 500 and 2000; test length is 10, 20 and 30; percentage of items showing DIF is 10% and 20%?

2. METHOD

This study compares five DIF detection methods using simulation, considering their power and Type I error rates. The model of the research is basic research since it is a research that will contribute to the previous knowledge in the literature by providing information about the performances of MH, LR, LORD, RAJU, and CSIBTEST methods (Karasar, 2021).

These DIF methods can demonstrate different conclusions according to different variables (e.g. trait distribution differences, sample sizes, length of the test, ratio of items with DIF, model type, and DIF type). The procedures performed to examine these five DIF methods in this study are presented below.

2.1. Simulation Conditions

A Monte Carlo simulation is utilized to analyze the Type I error rates and power of five DIF detection methods by changing independent variables: the sample sizes for the focal and reference groups, the test length, and the proportion of items showing DIF.

Sample size: Sample size per group can affect DIF detection rates. Hidalgo et al. (2016) indicate that the sample sizes are 250 per group for small size and 1000 per group for large size, and these sample sizes reflect situations in practice. Kaya et al. (2015) state that the small sample size is 250 per group in simulation studies to investigate DIF. Güler and Penfield (2009) identify 200-250 individuals per group as the small sample size and 1000 individuals per group representing the large sample size. Jodoin and Gierl (2001) used 250 per group for small and 1000 per group for large sample sizes in their simulation study. In this study, the sample size was simulated at 250 and 1000 per group for small and large sample sizes, respectively.

Test length: Test length can also affect DIF detection rates. If the number of items increases, more reliable results and more precise estimation of ability can be obtained (Narayanan & Swaminathan, 1996). Herrera and Gomez (2008) simulated 10 items, Rockoff (2018) simulated 10, 20 and 40 items, Gao (2019) simulated 20 and 40, Lopez (2012) simulated 15 and 30 items;

Glas and Meijer (2003) and Uysal et al. (2019) used 30 items in their simulation study. In this study, the test length was set at 10, 20 and 30 items, considered short tests in the literature (Narayanan & Swaminathan, 1994).

The proportion of DIF items: The proportion of items exhibiting DIF can affect DIF detection rates similar to test length but in the opposite direction. When the proportion of DIF items increases, DIF detection rates are likely to decrease (Jodoin & Gierl, 2001; Narayanan & Swaminathan, 1996). Demars and Lau (2011) stated that the percentages of DIF items were generally no more than 30%. Narayanan and Swaminathan (1994) indicated that the proportion of items showing DIF was either 10% or 20% in the simulation studies conducted to determine the effect of the proportion of items with DIF. Apinyapibal et al. (2015), Gao (2019), Holmes Finch and French (2007), Jodoin and Gierl (2001), Narayanan and Swaminathan (1996) used 10% and 20% DIF items in their study. In this study, the proportion of DIF items was 10% and 20%. Although these rates have been discussed in other studies, the performances of these five methods considering 10% and 20% DIF items have not been examined previously. The trait distribution (normal distribution), the model type (3PL), and the DIF type (uniform) remain constant in this research even though they also affect DIF detection methods.

In the simulation, sample size (500, 2000), test length (10, 20, 30), and percentage of items showing DIF (10%, 20%) are considered as manipulated conditions, while uniform DIF and 3PL models were considered as fixed conditions.

2.2. Data Generation

Data were generated using the 3PL IRT model which considers the case of answering correctly by chance. Item parameters were obtained through the WinGen3 software (Han & Hambleton, 2014). Tests consisting of 10, 20 and 30 items were created using the distributions of the item parameters obtained from an administration of the TIMSS 2019 paper-based Mathematics Test, a real test application to generate the data. The slope and the location parameters were generated using normal distributions with means of 1.3 and 0.531 and standard deviations of 0.357 and 0.52, respectively; the guessing parameters were set at 0.20 for all items because this parameter is near the upper end of its typically observed range (Reise & Waller, 2002). Lopez (2012) states that guessing is a realistic possibility in many testing applications and it is difficult to interpret the manipulations involving c-parameters in the context of DIF studies. Since fixing the c-parameters reduces Monte Carlo noise, a constant value of 0.20 is used for the guessing parameters.

A normal distribution with a mean of 0 and a standard deviation of 1 was used to generate the ability parameters. The differences between the location parameters of focus and reference groups for DIF items were taken as 0.60. Uniform DIF was simulated by randomly determining items. Items with DIF were applied using WinGen3 thus, 1-0 data were obtained for the focus and reference groups.

The simulation design consisted of 12 DIF conditions in total, which combined three different test lengths, two different sample sizes, and two different proportions of DIF items. Under each condition, 100 replications were made because it is common to obtain stable results (Kim, 2010). Thus, a total of 1200 data was generated. DIF analyses were performed for each data set with the five DIF methods mentioned before.

2.3. Data Analysis

The distributions of the slope and the location parameters obtained from an administration of the TIMSS 2019 paper-based Mathematics Test, a real test application, were determined using ARENA Input Analyser program. The test included number, algebra, geometry, data and probability items and was applied to 8th grade students. According to the results, the slope and location parameters distributions were normal distributions with means of 1.3 and 0.531 and standard deviations of 0.357 and 0.52, respectively. According to these distributions, the data were generated by the software WinGen3.

The data were analyzed and the methods were compared using dichoDif in the difR package (Magis et al., 2022) for data analysis of the R statistical software (version 4.0.2, R Core Team, 2022). Type I error, and the power rates were used to compare the performances of the methods. Type I error is the decision that the item shows DIF even though it does not actually show DIF. The power is the decision that the item showing DIF is determined as having DIF due to the analysis. Methods with high power rates and low Type I error rates are preferred for determining whether or not an item has DIF. According to Bradley (1978; as cited in Hidalgo et al., 2016), the Type I error rate should be between 0.025 and 0.075. The power of methods should be at least .80 to be sufficient and this criterion is widely used in the literature (Atar, 2007).

To compare the performances of the methods, a one-way ANOVA (assumptions have been met as the data for each group have a normal distribution, and these distributions have the homogeneity of variance) was also used for each study criteria to facilitate interpretations. In addition, factorial ANOVA was used to examine the interaction effects of the factors. The statistical significance findings of the respective analyses and post hoc comparisons were examined.

3. RESULTS

This research compares the DIF detecting methods using Type I error and their power rates under various conditions. For these 1200 data, it was examined whether there were significant differences between the performances of the DIF detection methods. The results in each condition are shown in Table 1.

As seen in Table 1, Type I error rates for small sample sizes in all conditions by MH and LR methods range from .034 to .081 and generally are lower than .075 and higher than .025, while Type I error rates for large sample sizes range from .063 to .174. When all methods are compared according to sample size, it is seen that Type I error rates are higher for large sample sizes.

Table 1. *Type I Error Rate and Power Rate by Study Procedures.*

Sample Size (Reference/ Focal)	Test length	%DIF	Number of DIF items	MH		LR		CSIBTEST	
				Type I	Power	Type I	Power	Type I	Power
500 (250/250)	10	10	1	.051	**.680**	.060	**.620**	**.096**	**.710**
		20	2	**.081**	**.640**	**.081**	**.575**	**.134**	**.670**
	20	10	2	.052	**.625**	.058	**.620**	**.079**	**.620**
		20	4	.034	**.030**	.039	**.060**	.056	**.060**
	30	10	3	.037	**.390**	.048	**.417**	.056	**.357**
		20	6	.069	**.457**	.067	**.440**	.071	**.383**
2000 (1000/1000)	10	10	1	.072	1.00	.074	1.00	**.128**	1.00
		20	2	**.174**	.985	**.160**	.980	**.271**	.985
	20	10	2	.068	.965	.063	.975	**.093**	.955
		20	4	**.131**	.838	**.114**	.890	**.173**	.853
	30	10	3	**.079**	1.00	.073	1.00	**.093**	1.00
		20	6	**.117**	.885	**.100**	.915	**.150**	.885

Table 1. *Continues*

Sample Size (Reference/ Focal)	Test length	%DIF	Number of DIF items	LORD		RAJU	
				Type I	Power	Type I	Power
500 (250/250)	10	10	1	**.023**	**.440**	**.023**	**.440**
		20	2	.051	**.405**	.051	**.405**
	20	10	2	.049	**.390**	.049	**.390**
		20	4	**.024**	**.025**	**.024**	**.025**
	30	10	3	**.080**	**.273**	**.080**	**.273**
		20	6	.046	**.293**	.046	**.293**
2000 (1000/1000)	10	10	1	.064	.990	.064	.990
		20	2	**.146**	.980	**.146**	.980
	20	10	2	.051	.925	.051	.925
		20	4	**.313**	**.735**	**.313**	**.735**
	30	10	3	.072	.990	.072	.990
		20	6	**.137**	**.768**	**.137**	**.768**

Note. MH=Mantel-Haenszel, LR= Logistic Regression, CSIBTEST=crossing simultaneous item bias test, LORD=Lord's chi square (χ^2), RAJU=Raju's area measure.

Power rates of all methods for small sample sizes are lower than .80 in all conditions and power rates of MH, CSIBTEST and LR methods for large sample sizes are above .80 in all conditions. Power rates of LORD and RAJU for large sample sizes are acceptable values, generally more than .80. When all methods are compared according to sample size, it is seen that power rates are higher for large sample sizes. In addition, the Type I error increases, and the power rate decreases in all methods as the ratio of the item with DIF increases for large sample sizes.

The comparison of the methods depending on the sample size can be seen more clearly in Figure 1 and Figure 2.

Figure 1. *Type I error rates for sample size.*

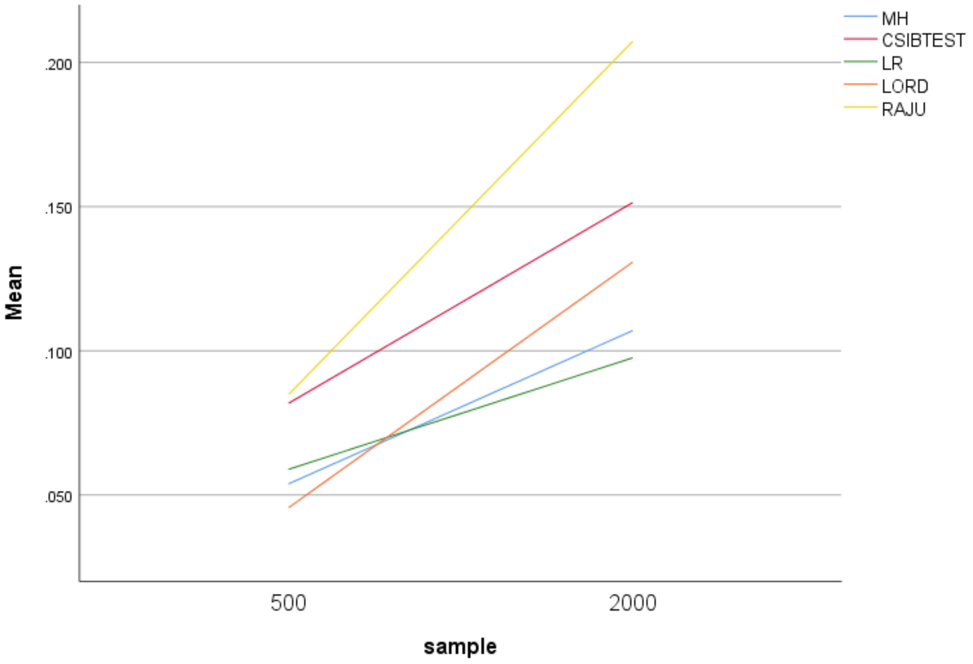
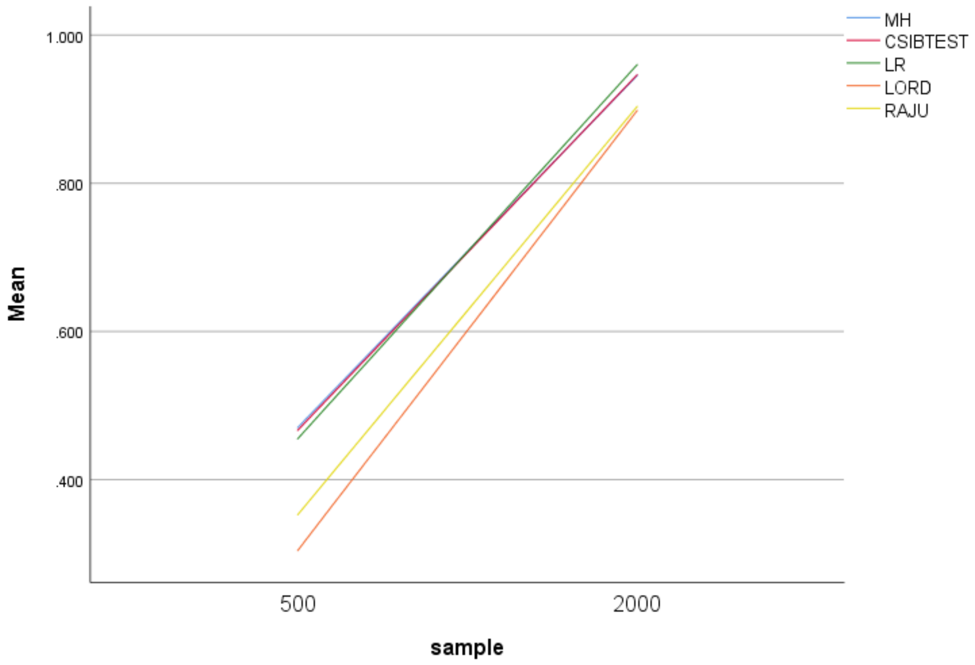


Figure 2. *Type I error rates for sample size.*



When Figure 1 and Figure 2 are examined, it can be seen more clearly that the Type I error increases and the power ratio decrease in all methods in the large sample than in the small sample. In addition, it is seen that RAJU has the highest Type I error, and MH, CSIBTEST and LR demonstrate significantly higher power rates than LORD and RAJU for both small and large sample sizes.

To facilitate interpretation, analyses of variance (ANOVA) for each procedure by manipulation were applied. The results for Type I error rates and power rates are shown in Tables 2 and 3, respectively.

It is found that the average Type I error rates of the methods are significantly different (F(4.5995) = 67.721, *p*<.05). Post hoc tests show that RAJU demonstrates significantly higher error rates (.146) than the other procedures. Then, it is found that CSIBTEST (.117) produces a significantly higher error rate than other methods except for RAJU. In addition, LR shows the lowest average Type I error rate, but there is no significant difference between MH, LR, and LORD.

Table 2. *ANOVA Results for Type I Error Rate by Study Procedures*

		MH		LR		CSIBTEST		LORD		RAJU	
	df	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
S	1	180.690*	.132	82.666*	.065	206.370*	.148	152.614*	.114	216.266*	.154
T	2	13.275*	.022	13.965*	.023	71.180*	.107	10.604*	.018	18.107*	.030
P	1	108.662*	.084	53.230*	.043	114.539*	.088	83.303*	.066	118.776*	.091
*S*T*	2	1.005	.002	1.456	.002	2.835	.005	20.727*	.034	26.279*	.042
*S*P*	1	45.015*	.037	31.174*	.026	74.646*	.059	113.937*	.088	113.508*	.087
*T*P*	2	10.675*	.018	7.200*	.012	16.633*	.027	19.044*	.031	15.160*	.025
*S*T*P*	2	8.785*	.015	5.494*	.009	4.601*	.008	27.182*	.044	46.651*	.073

Note. MH=Mantel-Haenszel, LR= Logistic Regression, CSIBTEST=crossing simultaneous item bias test, LORD=Lord's chi square (χ^2), RAJU=Raju's area measure. $^*p$<.05

As seen in Table 2, ANOVA results for Type I Error Rate for all procedures show that the main effects of sample size, test length, and proportion of DIF items are significant. Furthermore, significant *sample size x proportion of DIF items, test length x proportion of DIF items*, and *sample size x test length x proportion of DIF items* interactions are found for all methods. A significant *sample size x test length* interaction is found in LORD and RAJU, while it is not a significant interaction effect in other DIF detection methods.

Table 3. *ANOVA Results for Type I Error Rate by Study Procedures*

		MH		CSIBTEST		LR		LORD		RAJU	
	df	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2	*F*	η^2
S	1	1322.417*	.527	1314.089*	.525	1426.104*	.546	1903.149*	.616	1586.204*	.572
T	2	91.293*	.133	106.056*	.151	47.503*	.074	63.776*	.097	47.024*	.073
P	1	110.971*	.085	103.059*	.080	92.598*	.072	96.417*	.075	98.439*	.077
S*T	2	31.249*	.050	44.828*	.070	19.497*	.032	1.627	.003	.483	.001
S*P	1	15.721*	.013	18.436*	.015	23.855*	.020	.260	.000	.068	.000
T*P	2	73.230*	.110	55.534*	.085	52.614*	.081	30.717*	.049	62.915*	.096
S*T*P	2	53.690*	.083	45.536*	.071	43.585*	.068	20.045*	.033	34.049*	.054

Note. MH=Mantel-Haenszel, LR= Logistic Regression, CSIBTEST=crossing simultaneous item bias test, LORD=Lord's chi square (χ^2), RAJU=Raju's area measure. $^*p<.05$

It is found that the average power rates of the methods significantly differ, too (F(4.5995) = 22.298, *p*<.05). Post hoc tests show that MH (.708), CSIBTEST (.707), and LR (.708) demonstrate significantly higher power rates than LORD (.601) and RAJU (.628). There are no significant power rate differences between MH, CSIBTEST and LR, and between LORD and RAJU. As seen in Table 3, ANOVA results for power rate for all methods are affected by sample size, test length, and the proportion of items exhibiting DIF. In addition, *sample size x test length* and *sample size x proportion of DIF items* are found to be statistically significant for the MH, CSIBTEST, and LR methods, while all other interactions are found to be statistically significant for all methods. The significant *test length x proportion of DIF items* and *sample size x test length x proportion of DIF items* interactions are found for all methods.

When the main factors are examined by independent samples t-test and one-way ANOVA, Type I errors and power rates of all methods for large samples are significantly higher than Type I errors and power rates for small samples. Type I errors and power rates of all methods for the shortest test length (10 items) are significantly higher than Type I errors and power rates for others (20 and 30 items), except LORD and RAJU for Type I error rates. For these methods, they are significantly lower than others. However, there are no significant differences for Type I error rates and power rates in all methods between 20 and 30 items, except RAJU (The Type I error rate of 20 items is higher than 30 items). There is a significant Type I error rate difference for RAJU and a significant power level difference for MH between 20 and 30 items. Type I errors and power rates of all methods for 20% DIF items are significantly higher than those for 10% DIF items, except MH and CSIBTEST. There is no significant difference between 10% and 20% for them.

4. DISCUSSION and CONCLUSION

The existence of differential item functioning indicates that some situations need attention in a test. If items show DIF in a test, it indicates that different undesirable factors may affect the feature that the test intends to measure (Shealy & Stout, 1993). Therefore, it is important to identify procedures that can effectively detect DIF.

This study examines the efficacy of five DIF determination methods; MH, LR, LORD, RAJU, and CSIBTEST, considering various conditions. For this purpose, a simulation study was conducted considering real data parameters from an administration of the TIMSS 2019 paper-based Mathematics Test.

According to the results, it is found that the Type I error is low for the MH method and it gives acceptable results under many conditions (Marañón et al., 1997; Shealy & Stout, 1993). Guilera et al. (2013) discussed the Type I error and power of the MH method using the meta-analysis technique and found similar results for the MH method to this study. LR demonstrates the lowest average Type I error rate, and methods show slightly greater error rates than the nominal .075 error rate. These findings support the results of the study by Lopez (2012), which compared the efficacy of CSIBTEST, IRT-LR, and LR. LR had the lowest average Type I error rate, and CSIBTEST and IRT-LR demonstrated error rates that were greater than the nominal .05 level (Lopez, 2012). In addition, no significant differences between LR, MH, and LORD according to Type I error rate are found in this study. These findings are consistent with similar studies in the literature (DeMars, 2009; Erdem Keklik, 2014; Gierl et al., 2000; Rogers & Swaminathan, 1993; Uyar, 2015; Vaughn & Wang, 2010). According to Type I error and power rate, MH and LR have the lowest Type I error rate and the highest power rate. This finding supports the results of the research of Erdem Keklik (2014), which found that the MH and LR methods were similar and had lower Type I errors than IRT-LR when the trait distributions are normally distributed. It can be concluded that MH and LR are more sensitive to detecting items with DIF than other methods in this study.

When the methods are examined under different conditions, it is seen that their Type I errors, and power rates can differ according to the conditions. Swaminathan and Rogers (1990) indicated that the sample size affects the power of DIF detection procedures. In this study, when the small sample is compared with the large sample, it is seen that the Type I error and power ratios are higher in the large sample. Contrary to Holmes Finch and French (2007), these findings are in agreement with DeMars (2009), Güler and Penfield (2009), Li et al. (2012), and Roussos and Stout (1996).

It is expected that longer tests are likely to show more reliable scores. The power of the DIF methods is likely to increase with increasing test lengths (Narayanan & Swaminathan, 1996). However, Guilera et al. (2013) demonstrated that MH for tests with lengths from 20 to 40 items showed lower Type I error and power than shorter tests. In this study, Type I errors and power rates of tests with 20 and 30 items are found to be significantly lower than the shorter test (10 items), which is consistent with Guilera et al. (2013), Kim (2010), Lopez (2012) and Uttaro and Millsap (1994).

Fidalgo et al. (2000) stated that the greater the number of items with DIF, the greater the Type I error. The finding that the Type I error and power rates of all methods increase as the ratio of the item with DIF increases is consistent with the results in the literature (Atalay Kabasakal et al., 2014; Finch, 2005; Guilera et al., 2013; Holmes Finch & French, 2007; Uyar, 2015).

When the interaction effects are examined, it is seen that Type I errors and the power rates differ according to the interactions of *test length x proportion of DIF items* and *sample size x test length x proportion of DIF items* for all methods. Type I errors, and the power rates differ according to *sample size x test length* interactions of the LORD and RAJU. Type I errors differ according to the interactions of *sample size x proportion of DIF items* for all methods, while the power rates differ for only MH, CSIBTEST, and LR. It can be concluded that the interaction effect of the variables can differentiate the Type I errors and power ratios of the methods. Thus, it is thought that interaction effects should be taken into account when using the methods.

To sum up, when the results obtained from this study and other relevant research results are evaluated together, LR and MH are used as a reason for preference, especially in small samples, as they have the lowest Type I error and the highest power rate in detecting uniform DIF. It is seen that Type I errors, and power rates of the methods can differ according to the conditions. So, the preferred DIF determination methods should be chosen considering the applied situations and requirements of the theories (e.g. IRT-based DIF detection methods require a large sample size). It can be stated that which DIF methods to use should be decided by considering the conditions. As stated by Kelecioğlu et al. (2014) and Ayva Yörü and Atar (2019), at least two different DIF detection methods are suggested to be used to improve the reliability of the results, as different methods are seen to provide different results in certain situations. The methods to be used can be selected based on the properties of the application, such as sample size, test length etc. (e.g. LR and MH can be used if the sample size is small).

This study examined the efficacy of MH, LR, LORD, RAJU, and CSIBTEST methods considering various conditions. These DIF detection methods and used conditions are limitations of the study. Further studies may compare other DIF detection procedures based on CTT and IRT and the differences between them can be analysed according to Type I errors and their power rates. The sample sizes were 250 and 1000 per group and the proportions of DIF items were 10% and 20%. Different sample sizes and ratios of items with DIF can be used. It would be better comparing especially 10% to 30%. Test lengths were taken 10, 20 and 30 as short test lengths. Short and long test lengths can be also used. The normal trait distribution, the 3PL model type, and the uniform DIF type remain constant. Further studies may examine non-uniform DIF with these procedures by changing the values of slope and location parameters. In addition, different trait distributions and model types (1PL or 2PL) can be researched in the future.

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

Münevver Başman https://orcid.org/0000-0003-3572-7982

REFERENCES

Apinyapibal, S., Lawthong, N., & Kanjanawasee, S. (2015). A comparative analysis of the efficacy of differential item functioning detection for dichotomously scored items among logistic regression, SIBTEST and raschtree methods. *Procedia-Social and Behavioral Sciences*, *191*, 21-25. https://doi.org/10.1016/j.sbspro.2015.04.664

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (type I error and power) of IRT likelihood ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning, *Educational Sciences: Theory and Practice, 14*(6), 2175-2193. https://doi.org/10.12738/estp.2014.6.2165

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures* [Unpublished doctoral dissertation]. University of Florida State.

Ayva Yörü, F.G., & Atar, H.Y. (2019). Determination of differential item functioning (DIF) according to SIBTEST, Lord's [Chi-squared], Raju's area measurement and Breslow-Day Methods. *Journal of Pedagogical Research*, *3*(3), 139-150. https://doi.org/10.33902/jpr.v3i3.137

Camilli, G., & Shepard, L.A. (1994*). Methods for identifying biased test items*. Sage Publications.

Camilli, G. (2006). Test fairness. In R.L. Brennan (Ed), *Educational Measurement* (4th ed., pp. 221–257). Rowman & Littlefield.

De Ayala, R.J. (2009). *The theory and practice of item response theory*. The Guilford Press.

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*, 149-170. https://doi.org/10.3102/1076998607313923

DeMars, C.E., & Lau, A. (2011). Differential item functioning detection with latent classes: how accurately can we detect who is responding differentially?. *Educational and Psychological Measurement*, *71*(4), 597-616. https://doi.org/10.1177/0013164411404221

Dorans, N.J., & Holland, P.W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, *1992*(1), i-40. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Embretson, S.E., & Reise, S.T. (2000). *Item response theory for psychologists*. Lawrance Erlbaum Associates.

Erdem Keklik, D. (2014). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel ve lojistik regresyon tekniklerinin karşılaştırılması [Comparison of Mantel-Haenszel and logistic regression techniques in detecting differential item functioning]. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(2), 12-25. https://doi.org/10.21031/epod.71099

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.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295. https://doi.org/10.1177/0146621605275728

Gao, X. (2019). *A comparison of six DIF detection methods* [Unpublished master's thesis]. University of Connecticut.

Gierl, M.J., Jodoin, M.G., & Ackerman, T.A. (2000, April 24-27). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large* [Paper presentation] In Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA, United States.

Glas, C.A., & Meijer, R.R. (2003). A Bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement*, *27*(3), 217-233. https://doi.org/10.1177/0146621603027003003

Guilera, G., Gomez-Benito, J., Hidalgo, M.D. & Sanchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods, 18*(4), 553-71. https://doi.org/10.1037/a0034306

Güler, N., & Penfield, R.D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314-329. https://doi.org/10.1111/j.1745-3984.2009.00083.x

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage.

Hambleton, R.K., Clauser, B.E., Mazor, K.M., & Jones, R.W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Han, K.T., & Hambleton, R.K. (2014). User's manual for WinGen3: Windows software that generates IRT model parameters and item responses (*Center for Educational Assessment Report No. 642)*. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Herrera, A., & Gomez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity, 42*(6), 739-755. https://doi.org/10.1007/s11135-006-9065-z

Hidalgo, M.D., López-Martínez, M.D., Gómez-Benito, J., & Guilera, G. (2016). A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests. *Psicothema*, *28*(1), 83-88. https://doi.org/10.7334/psicothema2015.142

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum

Holmes Finch, W., & 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

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education*, *14*(4), 329-349. https://doi.org/10.1207/S15324818AME1404_2

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17– 64). Rowman & Littlefield.

Karasar, N. (2021). *Bilimsel araştırma yöntemleri* [Scientific research methods]. Nobel Yayınları.

Kaya, Y., Leite, W., & Miller, M.D. (2015). A comparison of logistic regression models for DIF detection in polytomous items: the effect of small sample sizes and non-normality of ability distributions. *International Journal of Assessment Tools in Education*, *2*(1), 22-39. https://doi.org/10.21449/ijate.239563

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Seviye belirleme sınavı'nın madde yanlılığı açısından incelenmesi [Investigation of placement test in terms of item biasness]. *Elementary Education Online*, *13*(3), 934-953.

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*. Dissertation, Georgia State University.

Li, Y., Brooks, G.P., & Johanson, G.A. (2012). Item discrimination and Type I error in the detection of differential item functioning. *Educational and Psychological Measurement*, *72*(5), 847-861. https://doi.org/10.1177/0013164411432333

Li, H.H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647-677. https://doi.org/10.1007/BF02294041

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th Ed.). Upper Saddle River.

Lopez, G.E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-likelihood ratio test, crossing-SIBTEST, and logistic regression procedures* [Unpublished doctoral dissertation]. University of South Florida.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Magis, D., Beland, S., &Raiche, G. (2022). Collection of methods to detect dichotomous differential item functioning (DIF). Package 'difR'.

Marañón, P.P., Garcia, M.I.B., & Costas, C.S.L. (1997). Identification of nonuniform differential item functioning: A comparison of Mantel-Haenszel and item response theory analysis procedures. *Educational and Psychological Measurement*, *57*(4), 559-568. https://doi.org/10.1177/0013164497057004002

Mellenbergh, G.J. (1983). Conditional item bias methods. In S.H. Irvine & J.W. Berry (Eds.), *Human assessment and cultural factors* (pp. 293-302). Springer.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-103). MacMillan.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*(4), 315-328. https://doi.org/10.1177/014662169401800403

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show non-uniform DIF. *Applied Psychological Measurement, 20*(3), 257-274. https://doi.org/10.1177/014662169602000306

Oshima, T.C., & Morris, S.B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43-50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

Osterlind, S.J., & Everson, H.T. (2009). *Differential Item Functioning*. Sage.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. http://www.R-project.org/

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502. https://doi.org/10.1007/BF02294403

Reise, S.P., & Waller, N.G. (2002). Item response theory for dichotomous assessment data. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 88–122). Jossey-Bass.

Rockoff, D. (2018). *A randomization test for the detection of differential item functioning* [Unpublished doctoral dissertation]. The University of Arizona.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 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*, 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194. https://doi.org/10.1007/BF02294572

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*(1), 15–25. https://doi.org/10.1177/014662169401800102

Uyar, Ş. (2015). Gözlenen gruplara ve örtük sınıflara göre belirlenen değişen madde fonksiyonunun karşılaştırılması [Comparing differential item functioning based on manifest groups and latent classes] [Unpublished doctoral dissertation]. University of Hacettepe.

Uysal, İ., Ertuna, L., Ertaş, F.G. & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 133-148. https://doi.org/10.21031/epod.534312

Vaughn, B.K., & Wang, Q. (2010). DIF trees: using classifications trees to detect differential item functioning. *Educational and Psychological Measurement, 70*(6) 941–952. https://doi.org/10.1177/0013164410379326

Zumbo, B.D.A. (1999). *Handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and likert type item scores*. Ottowa.

International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 160–174

https://doi.org/10.21449/ijate.1094589

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

Reconfiguring assessment practices and strategies in online education during the pandemic

Jason V. Chavez [1,*], **Daisy D. Lamorinas** [2]

[1]Zamboanga Peninsula Polytechnic State University, Faculty of School of Business Administration, Philippines
[2]Western Mindanao State University, Faculty of WMSU Curuan Campus, Department of Filipino, Philippines

ARTICLE HISTORY

Received: Mar. 28, 2022
Accepted: Mar. 19, 2023

Keywords:
COVID-19 Pandemic
Assessment,
Practices,
Strategies,
Online Education.

Abstract: The rise of the COVID-19 pandemic has led to the overhaul of the conduct of teaching and learning particularly in the assessment of learners during a time of crisis trapped in many structural and practical challenges. This study examines the assessment practices and strategies to protect its quality and integrity in the delivery of teaching and learning among higher education students at Zamboanga Peninsula Polytechnic State University, Zamboanga City, Philippines. This research employs a comprehensive and reliable survey questionnaire on the assessment practices and strategies for assessment, including its quality and integrity. A total of 300 students and teachers were purposefully selected for the study. Based on the findings, practical assessment and skill assessment were among the most widely employed strategies by the teachers. There was a need for skill development in distance learning which calls teachers to integrate it into skill assessment strategies. The study yields the current practices of the teachers in assessing the academic performances of the students, strategies to execute their assessment practices that comply with the health protocols, and strategies to safeguard the quality and integrity of these assessments despite the difficulties in the learning environment. This study is integral to extending the body of knowledge regarding the different assessment practices and strategies and how these influence the delivery of online education. Nevertheless, academic institutions should reconfigure their assessment practices in terms of which of these suits well their stakeholders.

1. INTRODUCTION

The COVID-19 pandemic spread across the globe restricting numerous sectors from working intact, especially for higher education causing the closure of traditional classes. Because of the vulnerability of face-to-face classes, thousands of school closures had been implemented to curb the continuous increase in cases (Toquero, 2020). This closure affected more than 1.2 billion learners worldwide (Tria, 2020) while in the Philippines the current education shifted to online and modular access which also affected more than 28 million learners in the country (UNESCO, 2020).

The focus of this current study is to examine the assessment practices and strategies to protect its quality and integrity in the delivery of teaching and learning among higher education

*CONTACT: Jason V. Chavez ✉ jasonsight@yahoo.com 🖃 Zamboanga Peninsula Polytechnic State University, Faculty of School of Business Administration, Philippines

e-ISSN: 2148-7456 /© **IJATE** 2023

students at Zamboanga Peninsula Polytechnic State University during the crisis. The main choices of the educational sector in continuing the delivery of lessons are online and modular methods. Sharp and Sharp (2016) suggest that online instructors have to secure a learning experience that helps learners receive, preserve, and develop their predetermined skills because online learning is essentially "learning by them." Additionally, the inability to carry on through an online course may cause deterrence from registering for an online course in the future which increases the dropout and lessens the enrollment possibilities (Muljana & Luo, 2019).

However, as HEIs are becoming involved in online education, preserving honesty and integrity in this learning environment is significantly difficult to obtain (Cole & Swartz, 2013). In such a sense, more than 50% of the students are suspected to cheat on their final exams (Cole & Swartz, 2013), while the number might be more alarming to those institutions that do not have rules for academic dishonesty.

Major responsibilities of an academic institution are to connect students to their lessons through giving assignments and train teachers to be interactive with their students (Bailey, 2015). Conventionally, the assessment is the sole responsibility of the instructor and relies mostly on summative assessment methods (Sharp & Sharp, 2016). The method of assessment involves gathering information from an array of sources to develop a "rich and meaningful understanding" of student learning and to provide the essential information to improve future educational processes (Adzima, 2020). Furthermore, Sharp and Sharp (2016) and Adzima (2020) highlighted that assessment methods used in online learning environments depend mostly on learners' writing skills and the prominent concern among academic officials has often focused on the quality of educational experiences within an online class. Similarly, Adzima (2020) affirms that the beginning of alternative assessments comes as the result of the frustration of teachers because of the limitations of some conventional evaluation methods. Thus, it is difficult to differentiate between learners' performance from the course content and learners' writing skills; however, it is also interesting that more traditional educators are using alternative assessment methods.

As one of the biggest state colleges and universities in the Western Mindanao region of the Philippines, Zamboanga Peninsula Polytechnic State University (ZPPSU) made its curriculum aligned to the delivery of education according to the needs of students amidst the crisis. The institution implemented several restrictions to deal with the pandemic without compromising the quality of education, along with the incorporation of several assessment strategies to effectively assess the learning of students.

There is therefore a need to study the different strategies that teachers implement in online learning modality. Previous studies have modeled these strategies to be used such as socioeconomic inclusive (Fung et al., 2022), strict tracking (Heisig & Matthewes, 2022), project-based assessment (Beneroso, & Robinson, 2022) and portfolio assessment (Sanjaya, 2022), among others. Different studies have been conducted on these strategies but there is no research that reconfigures these assessment strategies as independent methods. Additionally, the literature is not able to determine which of these strategies are effective based on the demographic profiles of the stakeholders.

The rapid growth of technology is helping online learning to expand in enrollment, especially during this challenging time for the education sector. Technology brings students, from different locations together to interact, collaborate, and build a learning community (Muljana & Luo, 2019). During the crisis, technology bridges the gap between the students to come along the implementation of cyberspace learning because of this forcing situation. Activities make students "experience a sense of satisfaction, accomplishment, pride, and sometimes delight" (Bailey, 2015, p. 114) while building students' perseverance and sense of responsibility for the tasks assigned to meet the standards. Improved manifestations, the logic of autonomy, and the aptitude to focus or control learning are among the frequently cited benefits of self and peer assessment (Mao & Peck, 2013). However, the implementation of online learning introduces different risks and challenges to both the teachers and students, especially in higher education institutions (HEIs) (Tria, 2020) including academic dishonesty and forced cheating. Assessment of activities based on the perceptions of teachers influences the effective strategies while involving students in assessments, which includes modeling or communication regarding assessment processes (Mao & Peck, 2013).

Objectives. The present study focuses on the quality of online education being given to students. It seeks to determine the types of assessments tools teachers used for online education during a pandemic, identify the strategies used to carry out the assessment in online education, and find out the strategies employed to protect the quality and integrity of the assessment of online education during the Pandemic, and subsequently assess responses based on respondents' demographic profiles.

2. METHOD

2.1. Population and Samples

The study used purposive and convenience sampling constituting the college students from Zamboanga Peninsula Polytechnic State University who were currently enrolled for the academic year 2020-2021. The sample consisted of 200 students from different demographics and 100 instructors teaching at the university.

2.2. Research Instrument

The study was quantitative research following the survey descriptive-comparative approach, applicable in comparing the means of the variables. In this study, comparing the variables (e.g., gender, computer literacy, status, academic roles) was essential in determining which assessment tools and strategies were applied to a certain profile, hence, allowing academic institutions to employ such reconfiguration based on the profiles of their stakeholders.

There were three sets of original surveys to gather information on three categories, namely an assessment practices survey, a survey on assessment strategies for leniency and flexibility, and a survey on quality and integrity. Three experts on educational assessments were sought to validate the statements and content of the instruments. Additionally, before the actual collection of data and analysis, the researcher ran a validity test resulting in 0.94, 0.93, and 0.89 of Cronbach's alpha. This showed that the three sets of original survey questionnaires used in this study were credible and had internal consistency.

2.3. Collection of Data

The researcher secured permission and clearance from the academic head before the administration of the survey questionnaires. All the participants were furnished a copy of the approved letter to conduct research including its purpose, ethical conduct, and voluntary clause to take part in this research. Upon the agreement between the authorities and the researcher, online forms were used to facilitate the administration and collection of information from the respondents. The entire study lasted from October 2020 to September 2021. The questionnaires were administered to the respondents and retrieved about two weeks later in July 2021. The retrieval rate was 100%. The entire survey happened online, and no face-to-face interaction was done to follow the guidelines of the Department of Health Philippines.

2.4. Data Analysis

Data gathered were analyzed in terms of frequency distribution, Mean, Standard Deviation for all descriptive data. Independent *t*-test and ANOVA were used for the significant differences of the responses based on the respondents' profiles. This study sought to determine the differences of employed assessment strategies and tools based on the demographics of the participants. Hence, such parametric tests T-test and ANOVA were applicable to determine which group differed considering this type of analysis were hereby essential in reconfiguring the strategies to be implemented. Comparing the means by these parametric tests helped in identifying which tools or strategies were applicable to certain demographics.

2.5. Ethical Considerations

This study employed ethical standards to ensure protection, security, and safety of the participants. The methods of collecting data for this study were reviewed accordingly. It was ensured that all the participants of the study understood the purpose of conducting this research. Responses were kept confidential, and no third-party people had access to the data gathered. Only the researchers had the contact details and information of the participants.

3. RESULT

Question 1: What is the profile of the respondents in terms of: [i] Gender, [ii] Indigent Status, and [iii] Literacy?

Table 1. *Demographics of the Respondents.*

Category		Frequency (N)	Percentage
Gender	Male	110	30.0%
	Female	190	70.0%
Indigent Status	Indigent	168	59.0%
	Non-indigent	132	41.0%
Computer Literacy	Needs Training	164	57.0%
	Average to Advanced	136	43.0%
Academic Roles	Students	200	66.7%
	Teachers	100	33.3%

Table 1 presents different demographics corresponding to the categories under certain groups. Gender is divided into two groups by male that consisted of 110 (30%) respondents while 190 (70%) for females. Indigent status has 168 (59%) under indigent group and 132 (41%) respondents for non-indigent. Another demographic being presented is the Literacy Level, where 164 (57%) need training and 136 (43%) respondents are either average or advanced. Academic role is dominated by 200 students and 100 for teachers.

Most of the respondents are female. It is also remarkable that the respondents have insufficient technical skills and classify themselves as indigent.

Question 2: What are the types of assessments tools teachers use for online education during pandemic?

Table 2. *Assessment tools based on Gender.*

Gender	Assessment Tools	Mean	Remarks
Male	Portfolio-Based Assessment	3.30	High
Female		3.24	High
Male	Practical Assessment	3.16	High
Female		3.18	High
Male	Skill Assessment	3.23	High
Female		3.22	High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate, 2.81-3.40 high, 3.40-4.00 very high

Table 3. *Assessment tools based on Indigent.*

Indigent	Assessment Tools	Mean	Remarks
Indigent	Portfolio-Based Assessment	3.26	High
Non-indigent		3.32	High
Indigent	Practical Assessment	3.20	High
Non-indigent		3.25	High
Indigent	Skill Assessment	3.24	High
Non-indigent		3.28	High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate, 2.81-3.40 high, 3.40-4.00 very high

Table 4. *Assessment tools based on Literacy.*

Computer Literacy	Assessment Tools	Mean	Remarks
Needs Training	Portfolio-Based Assessment	3.20	High
Average to Advance		3.32	High
Needs Training	Practical Assessment	3.16	High
Average to Advance		3.19	High
Needs Training	Skill Assessment	3.15	High
Average to Advance		3.33	High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 5. *Assessment tools based on Academic Roles.*

Academic Roles	Assessment Tools	Mean	Remarks
Students	Portfolio-Based Assessment	3.26	High
Teachers		2.95	High
Students	Practical Assessment	3.16	High
Teachers		3.56	Very High
Students	Skill Assessment	3.23	High
Teachers		3.58	Very High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 6. *Assessment tools based on Overall Mean.*

Assessment Tools	Mean	Remarks
Portfolio-Based Assessment	3.23	High
Practical Assessment	3.23	High
Skill Assessment	3.28	High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Indicated in Table 2, the use of Portfolio-based assessment is common among male respondents having the mean of 3.30 interpreted as high preference level. In contrast to male respondents, female has the mean of 3.24. Practical assessment is being used by females with the mean of 3.18 than that of males who have 3.16 mean rate. Male respondents assess their students through skill assessment indicated in the mean 3.23 similar to the females having 3.22 mean rate. Neither of the tools is below a moderate level.

Table 3 shows that most non-indigent teachers use portfolio-based assessment indicated in the mean 3.32; indigent teachers moderately prefer portfolio-based assessment based on the 3.26 mean. Non-indigent ones also use practical assessment as a tool with a moderate mean 3.25. Indigents have a mean of 3.20 for a practical assessment. Skill assessment receives the mean 3.28 for non-indigent while 3.24 for the indigent. It is visible that all of the respondents moderately prefer the tool though a bit common among non-indigent.

Table 4 indicates that portfolio-based assessment is used by average to advanced respondents with the mean 3.32; in contrast to those who need training having a mean of 3.20. Practical assessment is used also by average to advanced teachers with a mean rate of 3.19 compared to those with limited skills with a mean rate of 3.16. For Computer literate respondents, they opt for skills assessment in online education with a mean rate of 3.33. The respondents moderately prefer the assessment tools in general.

Table 5 indicates the assessment tools being dominantly used during their lessons. As shown, teachers mark the use of skill assessment as the widely preferred assessment tool among their colleagues. In contrast, the students believe that the Portfolios-based assessment is most likely used to assess them. This also reveal that students are more output-centered while teachers are particular to the demonstrations of the lessons.

Overall mean shows the preference level of each of the determined tools. As Table 6 presents, highest remark among others is Skill Assessment with a mean rate of 3.28, while both Portfolio-based and Practical assessments have the mean score of 3.24. Skill assessment is commonly used among teachers than of practical and skills assessments. The general response is at moderate preference level.

Performance assessment derived from traditional approaches includes portfolio assessment along with competencies and skills assessment (Oudkerk Pool et al., 2020). Such practices are also available in this study. For students in this study, Portfolio-based assessment allows them to collect data and information that serve as evidence to their performances. Similarly, Oudkerk Pool et al., (2020) elaborated that evidence-based assessment needs a demonstration of the application of the lessons rather than only knowing those. This is the reason why skill assessment is the most widely used approach. However, it is unclear how the students demonstrated the applications of their lessons when it comes to skill assessment which Oudkerk Pool et al., 2020 argue could be unsatisfactory.

Portfolio-based assessment and practical assessments have also their limitations especially during online and modular approaches. In such a sense, professional institutions such as the Australian Computer Society (2001) regard it as important with which students and practitioners can demonstrate their knowledge and their ability to continually update their skills (Mao & Peck, 2013). Even before the pandemic, traditional assessment methods do this badly as they are developed for discipline fields with a low rate of change of knowledge because of one-time usage (Mao & Peck, 2013). Additionally, academic dishonesty is most likely to happen in portfolio-based assessment which affects the performance-outcome aspect of online education.

Question 3: What are the strategies used to carry out the assessment in online education during the Pandemic?

Table 7. *Assessment Strategies based on Gender.*

Gender	Assessment Strategies	Mean	Remarks
Male	Leniency	3.26	High
Female		3.38	High
Male	Flexibility	3.23	High
Female		3.39	High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 8. *Assessment Strategies based on Indigent.*

Indigent	Assessment Strategies	Mean	Remarks
Indigent	Leniency	3.32	High
Non-indigent		3.38	High
Indigent	Flexibility	3.30	High
Non-indigent		3.49	High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 9. *Assessment Strategies based on Literacy.*

Literacy	Assessment Strategies	Mean	Remarks
Needs Training	Leniency	3.31	High
Average to Advanced		3.39	High
Needs Training	Flexibility	3.27	High
Average to Advanced		3.43	Very High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 10. *Assessment Strategies based on Academic Roles.*

Academic Roles	Assessment Strategies	Mean	Remarks
Students	Leniency	3.34	High
Teachers		3.50	Very High
Students	Flexibility	3.34	High
Teachers		3.47	Very High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 11. *Assessment tools based on Overall Mean.*

Assessment Strategies	Mean	Remarks
Leniency	3.36	High
Flexibility	3.37	High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 7 indicates the use of the assessment strategies according to gender. The results show female teachers usually use both leniency and flexibility than the male respondents do with a mean of 3.38 and 3.39, respectively.

Table 8 shows that non-indigent respondents (3.32) are using the approach of leniency to their lessons than the indigent (3.32). Similarly, the non-indigent is flexible in their lessons as compared to the indigent ones. It is remarkable that the use of flexibility is at a high preference level for non-indigent respondents.

Table 9 has results for the use of assessment strategies according to the literacy levels of the respondents. As presented, individuals having average to advanced computer literacy are using both leniency (3.39) and flexibility (3.43) more commonly than those who need training.

Table 10 indicates that teachers are being lenient (3.50) to their lessons and activities, which is also agreed by the students. Similarly, students have also preferred the flexibility aspect of the course where their teachers consider their choice of how, and in what aspect their lesson must focus on.

As presented in Table 11, the respondents frequently apply the aspect of flexibility in their classes. It is also described that the assessment also follows leniency. Both have nearly equal preference levels which determine their usage depending on the applicability.

Distance learning does not need frequent face-to-face interaction of teachers and students (Naidu, 2017). Because of its applicability during the pandemic, distance learning has become a standardized teaching approach since in distance learning students have access to learning opportunities, at when or what pace, including examinations while students are enabled to take those whenever given to them (Naidu, 2017). It is found in this study that flexibility incorporates the teachers' considerations to how and why students find difficulties in coping in their lessons. Being flexible is a great choice for the teachers since they also struggle in their delivery of lessons as well as in teaching their students.

Another aspect is the lenient approach where it is described in this study that "higher grading standards consistently lead to higher achievement" like the argument of Gershenson (2020). However, since their tests are longitudinal study, the result of this study differs in a long run, but the central concept is somehow comparable.

Question 4: What strategies are employed to protect the quality and integrity of the assessment of online education during the Pandemic?

Table 12. *Assessment Strategies for Integrity Based on Gender.*

Gender	Assessment Strategies for Integrity	Mean	Remarks
Male	Parallel Validation	3.39	High
Female		2.98	High
Male	Randomization	3.02	High
Female		3.14	High
Male	Strict Condition	2.96	High
Female		3.15	High
Male	Penalization	2.91	High
Female		3.03	High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate, 2.81-3.40 high, 3.40-4.00 very high

Table 13. *Assessment Strategies for Integrity based on Indigent.*

Indigent	Assessment Strategies for Integrity	Mean	Remarks
Indigent	Parallel Validation	3.05	High
Non-indigent		3.06	High
Indigent	Randomization	3.16	High
Non-indigent		3.13	High
Indigent	Strict Condition	3.11	High
Non-indigent		3.20	High
Indigent	Penalization	3.02	High
Non-indigent		3.08	High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80, moderate 2.81-3.40 high, 3.40-4.00 very high

Table 14. *Assessment Strategies for Integrity Based on Computer Literacy.*

Computer Literacy	Assessment Strategies for Integrity	Mean	Remarks
Needs Training	Parallel Validation	2.97	High
Average to Advanced		3.09	High
Needs Training	Randomization	3.05	High
Average to Advanced		3.18	High
Needs Training	Strict Condition	3.00	High
Average to Advanced		3.21	High
Needs Training	Penalization	2.91	High
Average to Advanced		3.10	High

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate, 2.81-3.40 high, 3.40-4.00 very high

Table 15. *Assessment Strategies for Integrity Based on Academic Roles.*

Academic Roles	Assessment Strategies for Integrity	Mean	Remarks
Students	Parallel Validation	3.02	High
Teachers		3.13	High
Students	Randomization	3.11	High
Teachers		3.50	Very High
Students	Strict Condition	3.09	High
Teachers		3.41	Very High
Students	Penalization	2.99	Moderate
Teachers		3.34	Moderate

Legend: 1.0-1.60 very low, 1.61-2.20 low, 2.21-2.80 moderate, 2.81-3.40 high, 3.40-4.00 very high

Table 16. *Assessment Strategies for Integrity based on Overall Mean.*

Assessment Strategies for Integrity	Mean	Remarks
Parallel Validation	3.07	High
Randomization	3.16	High
Strict Condition	3.14	High
Penalization	3.05	High

Legend: 1.0-1.60 very low, 1.61-2.20 low 2.21-2.80 moderate 2.81-3.40 high 3.40-4.00 very high

Table 12 indicates the results for the use of assessment strategies to protect the integrity of activities during lessons. Parallel Validation is used by the males with a moderate mean score of 3.39. Female teachers incorporate Strict Conditions (3.25) and Randomization (3.24) as strategies for protecting integrity. Penalization is less likely to be used among the strategies with a mean score of 2.91 for male teachers and 3.03 for female teachers.

Table 13 shows that non-indigent teachers use strict conditions as strategies for protecting integrity (3.20). While Randomization is commonly used by indigent teachers (3.16), both indigent and non-indigent teachers prefer penalization and parallel validation less as assessment strategies for integrity.

Table 14 presents the data for literacy as groups for assessment strategies in protecting integrity. Average to advanced teachers use Strict Conditions (3.21) and Randomization (3.18). The mean scores for teachers needing training indicate Randomization as the most used method (3.05), followed by Strict Conditions (3.00). Parallel Validation and Penalization are less preferred by both respondents.

Table 15 shows that the most used assessment strategy to protect the integrity for teachers is Randomization (3.50), where they essentially randomize and change the pace of the questionnaire to minimize the possibility of tapping to past lessons during exams. The students also find this as a crucial condition to be engaged in honesty. Strict condition is also a choice for teachers.

As shown in Table 16, both Randomization (3.16) and Strict Conditions (3.14) are the most preferred assessment strategies for academic dishonesty, which is followed by Parallel Validation with a mean score 3.07. The least preferred method in protecting integrity is Penalization.

As being suggested Lee-Post and Hapke (2017) faculty should also change assignments routinely, not to prevent cheating but also to keep them fresh and relevant. This shows that aside from being able to prevent academic dishonesty, randomization of test questionnaires could also help students in learning which is why the preference and usability level is high. In contrast, Holden et al., (2021) argued that the authority must follow strict guidelines and standards concerning the preparation of texts to have a plagiarism-free classrooms. Similarly, on the essay written by Jenifer Garret, practiced surveillance tactics (e.g., multiple tests, and random test questionnaires) in classrooms could influence the way teachers give assignments and exams. Additionally, Stephens et al., (2021) state such an approach also impacts the culture of academic dishonesty. Consequently, both randomization and strict conditions have higher usage possibilities during online learning.

Nevertheless, the teachers in this study were less likely to punish and penalize their students for being dishonest; this is also the least used strategy in protecting integrity. Though in agreement with Holden et al., (2021), by adopting such an approach to control plagiarism, there is no significant effect because of no intellectual, moral, or ethical growth.

Question 5: Are there any significant differences between the assessment tools and strategies from determined demographics?

Table 17. *Significant Differences: Demographic Profile (Significant at 0.05).*

Demographics		F	Sig.	Remark
Gender	Flexibility	6.780	0.011	Significant
	Randomization	6.767	0.004	Significant
	Strict Condition	8.494	0.011	Significant
	Penalization	4.171	0.044	Significant
Indigent	Strict Conditions	5.685	0.019	Significant
Computer Literacy	Strict Conditions	4.965	0.028	Significant
Academic Role	Portfolio-based `	7.502	0.007	Significant
	Practical Assessment	4.841	0.029	Significant
	Flexibility	5.379	0.022	Significant
	Strict Condition	10.33	0.002	Significant
	Penalization	4.749	0.031	Significant

Table 17 summarizes the parametric test for mean differences. The *p*-value is significant at 0.05. This further reveals the results where the commonly used tools and strategies have differences in usage among respondents. As shown, all of the assessment strategies for integrity are significant by the usage and preference levels. Flexibility in the assessment of lessons has also a significant difference.

Academic role widely differs on the perspectives they have; it is presented that this also varies on the usability and accessibility of the assessment tools and strategies, though.

4. DISCUSSION and CONCLUSION

Question 1: What is the profile of the respondents in terms of: [i] Gender, [ii] Indigent Status, [iii] Computer Literacy, and [iv] Academic Roles?

High participation rate comes from the female teachers and students (70.0%), which means most of the responses are based on the female perspective. In that sense, the results under the gender category are more likely according to the female teachers and students. The gender is therefore found to be a factor for the differences in the perspective as well as the approaches being delivered in assessing the students.

Furthermore, the indigent status of the respondents is considered as the factor because of its effect on the usability and accessibility to crucial resources for online learning. As provided earlier, 168 indigent individuals participated in this study.

Computer literacy is also an important aspect of online learning. This is where the delivery matters when there could be challenges that the participants encounter. These challenges would certainly affect their completion of the activities and assessment in online media (Kaewsaiha & Chanchalor, 2019; Naidu, 2017).

Since the teachers are responsible for the implementation of assessments, they have the utmost control of how they commonly use them. As shown, the number of the students is twice as those of the teachers, which represents that most of the data come from how commonly used the approach is than how many actually use them. The students look into how one approach is frequently used while teachers look at how one approach seems to benefit them in assessing their students.

Question 2: What are the types of assessments tools teachers use for online education during pandemic?

The assessment tool that is mostly used by the teachers during the course is Portfolio-based Assessment for both male and female teachers. All of the tools presented are applicable for computer literate teachers and non-indigent teachers. It is described that the teachers use different methods based on their skills, knowledge, and ability to deliver them fluently and effectively.

Portfolio-based assessment is mostly used by teachers because of its ease and practicality. According to Oudkerk Pool et al., (2020), portfolios provide an overview of students' performances and their development within the course of online and modular approaches. Additionally, this also connects to the possibility of improving the quality of education because of its high usability. The teachers have this approach like the traditional one, which made the tool useful and preferred because of the familiarity and accessibility to the resources. In the study of Mao and Peck (2013), the teachers also revealed that portfolios offer improvement in educational efficiency because it removes the need for a separate graduate assessment mechanism and minimizes documentation effort of students. For such a reason, assessment is the "responsibility" of the instructors and teachers, and portfolios appear to be "forced activities" (Bailey, 2015). Additionally, students view portfolios, in either means, as a widely used approach, even before the pandemic.

Another preferred assessment tool is the use of Practical Assessment where the teachers give activities to their students to assess their ability to apply what they have learned. This would range from the video presentation, reading comprehension, task-centered activities, and performances. This further demonstrates their lessons and reflects them in such a manner they benefit from what they have learned. For students, skill assessment is applicable in distance learning, but technical issues are imminent. As supported by Kaewsaiha and Chanchalor's (2019), some teachers believe that the quality of the works submitted to them in distance learning is less likely aligned to their instructions. Notably, this current study identifies that skill assessment is applicable in distance learning for both indigent and non-indigent teachers.

Skill Assessment is also used by the teachers although it seems no direct recognition from the students. This shows that skills assessment potentially does not assess the students at all due to the barriers distance learning has.

Question 3: What are the strategies used to carry out the assessment in online education during the Pandemic?

Strategies for assessing the students include Flexibility and Leniency which have varying degrees of usage based on the preferences of the teachers. It is described that males are more lenient to their scoring style while females are flexible to their lessons. Computer literate is most likely flexible but could also adapt the lenient style. It is visible that both strategies have nearly equal usage possibilities during the pandemic.

Many distance students often do not set out to complete a course and often withdraw because of personal issues (e. g., psychological or modality) that have less relationship to the quality of study program they are in (Naidu, 2017). In this scenario, teachers tend to be “flexible” to their students to continue the interaction between and among themselves.

The situation of today causes anxiety to the students as well to the teachers, which could be why the institution tries to be considerate to increase the confidence and be eager to finish their course. Naidu (2017) also argues that it is “incorrect to expect” that students can complete the activities in online learning without the supervision of their teachers. Because of the difference in the learning goals of students, likewise, schools and districts would do well to create grading standards to assess their students (Gershenson, 2020), and this is where the leniency comes across.

Flexibility is available in higher education to assess their students based on their skills and ability to complete the tasks; and because of the current situation, the teachers also tend to be lenient to increase the sense of achievement among their students. However, in a general manner, Flexibility is mostly used by institutions today more than the Leniency. This finding agrees to that of Bailey’ (2015) study where the institution has to have an interactive design where the stakeholders are enabled in the sharing of strategies, experiences, testing of ideas, and sharing of results. Being flexible creates a climate where students are given the chance to be selective of the lessons and approaches that suit their current knowledge and capacity to execute the activities.

Question 4: What strategies are employed to protect the quality and integrity of the assessment of online education during the Pandemic?

Randomization is consistent for being the mostly used method in protecting the integrity of activities. The sort ranges from where the teachers create questionnaires that are related to the lessons and then essentially “tweaking” them for the following exams. Another available strategy is implementing surveillance, or guidelines that the students have to follow. It is remarkable that Penalization is less likely to be a choice for the strategy. Similar to *Question 4*, the teachers prefer the Penalization approach less because they are lenient and flexible to their lessons.

The pandemic made online education challenging because of the presence of academic dishonesty in online media. There are many factors that cause the students to be engaged in the plagiarism culture. One of the methods used by the teachers is randomly tabulating the question which also showed positive results in minimizing academic dishonesty according to Cole and Swartz (2013). Randomization makes the questions appear different to assess the understanding of the students and lessen the possibility to plagiarize the exam based on the previous exam results.

Additionally, Strict Conditions are also used because these discourage the students to be dishonest and urge them to follow specific guidelines set by their teachers. This is effective to building the culture of academic honesty by following the instructions and directions (e. g., criteria, and scorecards). In such a manner, the implementation of strict conditions in each activity controls cheating and dishonesty.

Penalization is less likely to be a choice for the teachers and students to control the cases of dishonesty. In fact, this has a less known effect based on recent studies.

However, it is suggested by Lee-Post and Hapke (2017) that the conflict of academic dishonesty could be signified along with values and ethical development among students. Additionally, colleges should reassess their prominent approaches towards cheating and academic dishonesty. Likewise, in this study, if current approaches are not maintaining a satisfactory level of academic honesty, approaches might also follow new methods either determined by the institution or not.

Question 5: Are there any significant differences between the assessment tools and strategies from determined demographics?

In this study, males prefer leniency while females prefer flexibility. Randomization, Strict Condition, and Penalization are different by gender which means there are different mechanisms that teachers could utilize online education. The indigent status also differs in a strict climate because non-indigent teachers tend to use this approach more frequently than the others. Similarly, the strict condition is used mostly by the computer-literate teachers due to their ability to locate dishonesty with their technical skills.

In the study of Ching and Hsu (2015), females prefer audio/video discussion because it allows them to have effective communication. In this current study, females also prefer such type of an assessment strategy because of its efficacy in delivering their performances. Likewise, professors were also asked to be lenient to student's schedules and deliver their lessons in flexible mode (Singgih, 2021). In terms of having a strict condition and penalization, females tend to see this as an approach utilized by their teachers in online education. Previous studies were able to determine that male students have higher tendency to cheat in online settings (Adzima, 2020) which this study was able to determine why males do not see the strictness as feasible in online learning.

Limited knowledge on the use of computers can impact the teaching experiences of teachers. Teachers require cognitive skills (e.g., decrypt images) and procedural skills (e.g., processing files) which are essential when using computer programs (Liu et al., 2020). Instructors need to have such skills to combat cheating in online assessment (Gamage et al., 2020). Hence, this showed that the ability to use different detecting strategies during assessments requires higher computer literacy.

Gender roles have the crucial information that displays the difference in the responses. Portfolio-based and Practical Assessments are widely recognized among teachers and students. Flexibility is preferred over leniency. While Strict condition is constantly a common approach, the respondents are also eying for possible penalization where the two are complementary approaches. However, the overall data differ from the other perspective. For such conditions, Slade et al., (2022) suggested that higher education have to reassess their purpose of assessment if they want to equip their learners with crucial skills and competencies for future workplace.

Computer literacy is not a factor for the use of assessment strategies (e. g., leniency and flexibility). No factor had displayed differences for the assessment tools (e. g., portfolio-based, practical, and skill assessments).

4.1 Recommendations

Higher education must have a holistic approach in assessing their students to maximize the learning they obtain during the pandemic. It is significant to follow the preferences and the ability of stakeholders to certain assessment tools and strategies to have the effect be relevant and timely to the needs of the teachers and students. There are varying methods that are found to be effective at some sort but at least to the other; in this sense, understanding the actual situation of higher education in advancing to online learning could yield enormous benefits for the institution. Guidelines of preferred assessments practices can be integrated as a policy on similar situations in the future. The teachers should also be aware of what assessment tools and strategies are applicable to them to increase the capacities and skills of their students amidst the pandemic.

1. Academic institutions have to employ assessment tools that are widely applicable to their students and will protect the quality and integrity of the assessment specifically following the initial guidelines of CHED (2020 & 2021) during the pandemic.

2. Education departments need to develop a guidebook or manual and provide support and training for the conduct of lenient, flexible, and quality assessments to the students and to the teachers needed further competence about assessments during crisis.

3. Education departments should continuously provide feedback to the educators and institutions and study assessment practices while navigating the educational environments during the pandemic and even post-pandemic until the institutions can come up with reliable policies and guidelines on assessments.

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**: Zamboanga Peninsula Polytechnic State University, ZPPSU-REOC-2021-001, May 24, 2021

Authorship Contribution Statement

Jason V. Chavez: Investigation, Resources, Analysis based on the automated item selection procedure, and Writing-original draft. **Daisy D. Lamorinas**: Investigation, Resources, Analysis based on the automated item selection procedure, and Writing-original drafts.

Orcid

Jason V. Chavez https://orcid.org/0000-0002-4986-1034
Daisy D. Lamorinas https://orcid.org/0000-0002-1725-0626

REFERENCES

Adzima, K. (2020). Examining online cheating in higher education using traditional classroom cheating as a guide. *Electronic Journal of E-Learning*, *18*(6), 476-493.

Bailey, S., Hendricks, S., &Applewhite, S. (2015). Student perspectives of assessment strategies in online courses. *Journal of Interactive Online Learning*, *13*(3), 112-125.

Beneroso, D., & Robinson, J. (2022). Online project-based learning in engineering design: Supporting the acquisition of design skills. *Education for Chemical Engineers*, *38*(1), 38-47.

Ching, Y.H., & Hsu, Y.C. (2015). Online graduate students' preferences of discussion modality: Does gender matter? *Journal of Online Learning and Teaching*, *11*(1), 31-41.

Cole, M.T., & Swartz, L.B. (2013). Understanding academic integrity in the online learning environment: A survey of graduate and undergraduate business students. *ASBBS Proceedings*, *20*(1), 738.

Fung, C.Y., Su, S.I., Perry, E.J., & Garcia, M.B. (2022). *Development of a socioeconomic inclusive assessment framework for online learning in higher education*. In Socioeconomic inclusion during an era of online education (pp. 23-46). IGI Global.

Gamage, K.A., Silva, E.K.D., & Gunawardhana, N. (2020). Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Education Sciences*, *10*(11), 301.

Gershenson, S. (2020). *Great Expectations: The Impact of Rigorous Grading Practices on Student Achievement*. Thomas B. Fordham Institute, pp. 1-48.

Heisig, J.P., & Matthewes, S.H. (2022). No Evidence that Strict Educational Tracking Improves Student Performance through Classroom Homogeneity: A Critical Reanalysis of Esser and Seuring (2020). *Zeitschrift für Soziologie*, *51*(1), 99-111.

Holden, O.L., Norris, M.E., & Kuhlmeier, V.A. (2021). Academic integrity in online assessment: A research review. *In Frontiers in Education*, *6*(1), 1-13. https://doi.org/10.3389/feduc.2021.639814

Kaewsaiha, P., &Chanchalor, S. (2019). Survey on the use of learning management systems and online skill-based assessment in Thai teacher universities. *Education*, *100*(1), 92.

Lee-Post, A., & Hapke, H. (2017). Online learning integrity approaches: Current practices and future solutions. *Online Learning, 21*(1), 135-145.

Liu, Z.J., Tretyakova, N., Fedorov, V., & Kharakhordina, M. (2020). Digital literacy and digital didactics as the basis for new learning models development. *International Journal of Emerging Technologies in Learning, 15*(14), 4-18.

Mao, J., & Peck, K. (2013). Assessment strategies, self-regulated learning skills, and perceptions of assessment in online learning. *Quarterly Review of Distance Education, 14*(2), 75-95.

Muljana, P.S., &Luo, T. (2019). Factors contributing to student retention in online learning and recommended strategies for improvement: A systematic literature review. *Journal of Information Technology Education: Research, 18*(1), 19-57.

Naidu, S. (2017). Openness and flexibility are the norm, but what are the challenges?

Oudkerk Pool, A., Jaarsma, A.D.C., Driessen, E.W., & Govaerts, M.J. (2020). Student perspectives on competency-based portfolios: Does a portfolio reflect their competence development? *Perspectives on medical education, 9*(1), 166-172.

Sanjaya, D.B., Suartama, I.K., & Suastika, I.N. (2022). The Effect of the Conflict Resolution Learning Model and Portfolio Assessment on the Students' Learning Outcomes of Civic Education. *International Journal of Instruction, 15*(1), 473-488.

Sharp, L.A., & Sharp, J.H. (2016). Enhancing student success in online learning experiences through the use of self-regulation strategies. *Journal on Excellence in College Teaching, 27*(2), 57-75.

Singgih, I. (2021). *Being a 'Lenient'Math Professor*. In Proceedings of the 1st International Conference on Education, Humanities, Health and Agriculture, ICEHHA 2021, 3-4 June 2021, Ruteng, Flores, Indonesia.

Slade, C., Lawrie, G., Taptamat, N., Browne, E., Sheppard, K., & Matthews, K.E. (2022). Insights into how academics reframed their assessment during a pandemic: disciplinary variation and assessment as afterthought. *Assessment & Evaluation in Higher Education, 47*(4), 588-605.

Stephens, J.M., Watson, P.W.S.J., Alansari, M., Lee, G., & Turnbull, S.M. (2021). Can online academic integrity instruction affect university students' perceptions of and engagement in academic dishonesty? Results from a natural experiment in New Zealand. *Frontiers in Psychology, 12*(1), 1-16.

Teclehaimanot, B., You, J., Franz, D.R., Xiao, M., & Hochberg, S.A. (2018). Ensuring academic integrity in online courses: A case analysis in three testing environments. *The Quarterly Review of Distance Education, 12*(1), 47-52.

Toquero, C.M. (2020). Challenges and opportunities for higher education amid the COVID-19 pandemic: The Philippine context. *Pedagogical Research, 5*(4), 1-5.

Tria, J.Z. (2020). The COVID-19 pandemic through the lens of education in the Philippines: The new normal. *International Journal of Pedagogical Development and Lifelong Learning, 1*(1), 2-4.