Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *Güz* 2024 *Autumn* 2024 *Cilt: 15-Sayı: 3* *Volume: 15-Issue: 3*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

İletişim
e-posta: epodderdergi@gmail.com
Web: https://dergipark.org.tr/tr/pub/epod

Contact
e-mail: epodderdergi@gmail.com
Web: http://dergipark.org.tr/tr/pub/epod

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayımlanan hakemli uluslararası bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is a international refereed journal that is published four times a year. The responsibility lies with the authors of papers.

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi), EBSCO

Hakem Kurulu / Referee Board

Gökhan AKSU (Adnan Menderes Üni.)
Görkem CEYHAN (Muş Alparslan Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice Çiğdem BULUT (Northern Alberta IT)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Bolu Abant İzzet Baysal Üni.)
İbrahim YILDIRIM (Gaziantep Üni.)
İbrahim UYSAL (Bolu Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kadriye Belgin DEMİRUS (Başkent Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mahmut Sami KOYUNCU (Afyon Üni.)
Mahmut Sami YİĞİTER (Ankara Sosyal B. Üniv.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Merve ŞAHİN KÜRŞAD (TED Üni.)
Metin BULUŞ (Adıyaman Üni.)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özen YILDIRIM (Pamukkale Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)
Ragıp TERZİ (Harran Üni.)
Sedat ŞEN (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Safiye BİLİCAN DEMİR (Kocaeli Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Seher YALÇIN (Ankara Üni.)
Selen DEMİRTAŞ ZORBAZ (Ordu Üni.)
Selma ŞENEL (Balıkesir Üni.)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sait Çüm (MEB)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Sedat ŞEN (Harran Üni.)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serap BÜYÜKKIDIK (Sinop Üni.)
Serkan ARIKAN (Boğaziçi Üni.)
Seval KIZILDAĞ ŞAHİN (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem DEMİRKOL (Ordu Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Süleyman DEMİR (Sakarya Üni.)
Sümeyra SOYSAL (Necmettin Erbakan Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT (İzmir Demokrasi Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Wenchao MA (University of Alabama)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Yusuf KARA (Southern Methodist University)
Zekeriya NARTGÜN (Bolu Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni.)

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 183-192

The Comparison of PISA 2015-2018 Mathematics Trend Items Based on Item Response Times

Muhsin POLAT* Hülya KELECİOĞLU**

Abstract

This study aims to explore the intricate relationship between students' response times, item characteristics, and the effort invested during the Programme for International Student Assessment (PISA) 2015 and 2018 cycles. Through the analysis of data obtained from 69 mathematics trend items, administered in a computer-based format across both PISA 2015 and 2018 cycles with a focus on the Türkiye sample, this research investigates the dynamics of students' response times and their implications on effort and item characteristics. The findings reveal a significant increase in students' mean response times in the 2018 cycle compared to 2015, indicating potentially heightened effort and solution behavior. Notably, item formats exerted a substantial influence on response times, with open-ended items consistently eliciting longer response times compared to multiple-choice items. Additionally, a correlation between response times and item difficulty emerged, suggesting that more challenging items tend to consume more time, possibly due to the complexity of involved cognitive processes. Item-based effort, assessed through Response Time Fidelity (RTF) indices, highlighted that the majority of students exhibited solution behavior across both cycles to the items. Moreover, a decrease in the proportion of students displaying rapid-guessing behavior was observed in the 2018 cycle, potentially reflecting increased engagement with the assessment. While providing insights into the interplay of response times, item characteristics, and effort, this study emphasizes the need for further exploration into the multifaceted nature of effort in educational assessments. Overall, this research contributes valuable perspectives on nuances surrounding test performance and effort evaluation within PISA mathematics assessments.

Keywords: item response time, PISA, response-time effort, rapid-guessing

Introduction

In the Programme for International Student Assessment (PISA) and other tracking assessment procedures, where students' acquired knowledge and skills are evaluated, the expected behavior of students is to consciously apply their acquired knowledge and skills to respond to items. The scores obtained from these tests are used to assess individuals' knowledge and abilities in terms of what they know and what they can do. During these assessments, it is assumed that individuals engage in effortful attempts to answer the test items. In fact, this assumption is made across all measurement processes (Wise & Kong, 2005). An individual engages in an interaction with test items, exhibiting effort to respond to each item to the best of their ability. Test-taking effort is commonly characterized by a student's active involvement and dedication to resources aimed at achieving the most favorable outcome on the examination (Debeer et al., 2017). In high-stakes tests, individuals are expected to exhibit this behavior because the test scores hold significance for the individual. In placement exams for institutions, university entrance exams, course passing exams, and similar assessments, students consciously exhibit effort to respond to test items. However, in low-stakes national and international monitoring tests, students can respond to test items without exhibiting much effort and complete the test within a very short period of time. In this case, it can lead to variance that is unrelated to the structure that the test aims to measure (construct-irrelevant variance), and test scores may underestimate the examinee's true ability. Failing to account for the impact of effort exerted during testing could compromise the validity

* National Educational Expert, Republic of Türkiye Ministry of National Education, Ankara-Türkiye, muhsinpolat58@gmail.com, ORCID ID: 0009-0003-2897-3189
** Prof. Dr. Hacettepe University, Faculty of Education, Ankara- Türkiye, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Polat, M. & Kelecioğlu, H. (2024). The comparison of PISA 2015-2018 mathematics trend items based on item response times. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 183-192. https://doi.org/10.21031/epod.1398317

Received: 1.12.2023
Accepted: 24.09.2024

of test outcomes (Michaelides & Militsa, 2022). In cases of a lack of effort from individuals, two situations will emerge regarding test scores. The first situation is that when individuals do not exhibit sufficient effort, it will lead to a negative bias in the test scores. The second situation is that when individuals exhibit different levels of effort, it will result in variability in effort bias among individuals (Wise et al., 2006). Due to the construct-irrelevant variance caused by effort bias, the estimated levels of individual ability derived from these test scores will likely be lower than the actual ability levels. Individuals with high motivation and who exhibit effort in solving the test items will generally have higher test performance compared to those who show low effort (Eklöf et al., 2014; Rios & Guo, 2020). Therefore, in such a situation where variance unrelated to the construct occurs, the validity of evaluations made based on test scores and the decisions made will be low. Validity is necessary for interpreting test scores and using these scores for any purpose (AERA et al., 2014). The impact of factors not aligned with the test's objectives on test scores diminishes their validity, resulting in flawed deductions drawn from these scores.

The low effort that individuals put into answering test items directly impacts the validity of the test. Therefore, measuring individuals' efforts and revealing their impact is highly important. In the literature, efforts are measured using various methods. The first one is administering a self-report scale after the test for individuals to indicate how much effort they exerted. In this method, Likert-type scales are commonly used to reveal individuals' efforts. However, the objectivity of the information gathered about effort through self-reporting may be low (Wise & DeMars, 2006). Another method developed to measure individuals' efforts is person-fit statistics. Person-fit statistics aim to identify individuals' abnormal responses. For this purpose, each individual's response pattern is compared with measurement models (Meijer, 1996). Abnormal responses can include copied, careless, or random answers. Therefore, since abnormal responses don't solely indicate a lack of effort, using this statistic might not yield accurate results. Another method developed to identify responses without exhibiting effort, also used in this study, involves evaluating item response times. In this method, effort is defined by individuals' response times to items (Schnipke & Scrams, 1997; Wise & Kong, 2005). According to this method, individuals develop two types of behavior when encountering an item. First, individuals exhibit effort to answer the item correctly and attempt to solve it; this behavior is termed "solution behavior". Second, individuals do not contemplate the item and answer it without attempting to solve it; this behavior is termed "rapid-guessing behavior". Response times expended by individuals to solve the item are used to distinguish between these two behaviors (Wise & Kong, 2005).

Computerized tests enable the gathering of response times and diverse process-related data. Such tests make it possible to measure response time at the item level. The response time denotes the duration test-takers require to answer a specific item within the test context (Lee & Jia, 2014). Response time has been seen as a valid indicator of test-taking effort (Wise & Kong, 2005). As response time offers insights into examinee test-taking behavior on a per-item basis, it empowers researchers to monitor potential fluctuations in effort throughout the testing session (Wise & Kingsbury, 2016). For instance, item position within a test has been extensively examined as a significant factor influencing examinee effort; specifically, effort tends to decrease towards the end of a testing session (Debeer et al., 2014; Pools & Monseur, 2021). Item characteristics with less reading material and more answer options were associated with less rapid-guessing behavior (Setzer et al., 2013). DeMars (2000) indicated evidence for higher item non-response and lower effort in low-stakes constructed responses compared to multiple-choice items (DeMars, 2000). Therefore, examining response times across item types and characteristics between PISA cycles will enable the assessment of test-takers' efforts.

Wise and Kong (2005) developed the Response Time Effort (RTE) index to determine students' behavioral types toward items and their effort directed toward the test. This index utilizes the response time when students encounter an item to ascertain whether they exhibit solution behavior or engage in rapid-guessing. By assessing students' responses to all items on the test collectively, the Response Time Effort (RTE) index is constructed to represent the effort exhibited by students. The RTE indices range between 0 and 1, representing the proportion of solution behavior displayed during the test. As the RTE value approaches 1, it indicates that the student exhibited strong effort to solve the test, while a value closer to 0 suggests minimal effort was exerted. Wise (2006) similarly utilized item response times to

develop an index that indicates how much effort was exhibited on each item in the test. This index, named Response Time Fidelity (RTF), demonstrates the extent to which items are solved with solution behavior by students. The RTF index for items ranges between 0 and 1. As the index approaches 1, it indicates that a large number of students exhibited effort on the item, while a value closer to 0 suggests minimal effort from students. As a result, the RTE index reflects an individual student's effort across all items, while the RTF represents the collective effort of all students on a single item. The RTE index signifies effort pertaining to an individual, whereas the RTF indicates effort directed at a particular item. The normative threshold-setting methods were employed to generate RTE and RTF indices and determine individuals' behavioral types toward items. In these methods, the mean item response time for each item is calculated, and threshold points are established by taking a given percentage of these calculated item response times (Wise & Ma, 2012). In this study, thresholds were determined by taking the 10th and 20th percentiles of response times: NT10 and NT20 methods. It is believed that the outcome of the study will contribute to a better understanding of the results from the PISA assessment.

Item response time can be utilized for various purposes, such as item selection in computer-adaptive testing (Lee & Haberman, 2016), its relationship with student motivation (Wise & Kingsbury, 2016), detecting abnormal response behaviors (van der Linden & Guo, 2008), its association with test-taking behaviors (rapid-guessing or solution-based behavior) (Wise, 2006), and serving as an additional source of information to improve the accuracy of ability and item parameter estimations (Petscher et al., 2015; Wise & DeMars, 2006). Understanding the time and effort individuals spend on solving items is crucial for minimizing errors in item and ability parameter estimations, as well as reducing measurement error in test scores. Additionally, analyzing item response times contributes to a deeper understanding of the interaction between respondents and test items (Ju, 2021). This study offers a detailed comparison of item response times from the PISA 2015 and 2018 assessments, focusing on item format, item parameters, and response time fidelity differences. By examining these variations, the research provides insights into how different item types influence response behaviors, contributing to more accurate estimations of student ability and item performance. Furthermore, understanding these time differences helps improve the design of future assessments, ensuring that test validity and fairness are maintained. This analysis also adds to the existing literature by exploring the impact of test formats on response time dynamics in large-scale international assessments.

Purpose of the Study

Within this study, mean response times for trend mathematics items shared between the PISA 2015 and PISA 2018 cycles were compared based on item formats, assessment frameworks in which items were placed, and item parameters. Additionally, RTF indices for items were compared using the NT10 and NT20 methods for both assessment cycles.

Method

Data from 69 trend items in the mathematics subtest of the PISA 2015 and 2018 cycles, specific to Türkiye, were used in this study. Both cycles were conducted in a computer-based format, referred to as Computer Based Assessment (OECD, 2023). In this study, three variables are used: response time, response (coded and scored) and item characteristics (item difficulty, item discrimination and item format). These variables are available in PISA 2015 and 2018 public use datasets and technical reports. Response time, a continuous variable derived from the process data associated with each item, indicates the total duration spent by individual students on the respective items. The scored (by computer) and coded (by human) response variables consist of seven categories as follows: 0 = No Credit, 1 = Full Credit, 5=Valid Skip, 6 = Not Reached, 7 = Not Applicable, 8 = Invalid, and 9 = No Response (OECD, 2023). Categories 5, 6, 7, and 8 are recoded as missing values, and category 9 is recoded as 0, employing the identical coding rules for missing scores as delineated in PISA's methodology (OECD, 2023). Item difficulty is the proportion of correct responses for all items and item discrimination is the item-total correlation. These variables were used to examine the relationship between item characteristics and item response times, as well as to analyze the RTF indices for the items in both 2015 and 2018. Wise (2006) developed the SBij equation to measure Response Time Fidelity (RTF). Here, Ti represents the threshold point that delineates between rapid-guessing behavior and solution behavior for each item i,

while RTj denotes the response time for item i by individual j. The threshold points are determined based on normative threshold values developed by Wise and Kong (2006) using the NT10 and NT20 criteria. NT10 and NT20 cutoffs are obtained by calculating the mean response time for each item separately and taking ten and twenty percent of these calculated mean response times, respectively. Accordingly, the behavior of individual j toward item i is calculated as follows.

$$SB_{ij} = \begin{Bmatrix} 1\ if\ RT_{ij}\ \geq\ T_i \\ 0\ otherwise. \end{Bmatrix} \quad (1)$$

SBij represents the behavior of individual j toward item i and is binary. Accordingly, if an individual's response time for an item is greater than the cutoff point, the behavior is considered solution-oriented. If it is smaller, it's regarded as rapid-guessing behavior.

Response Time Fidelity (RTF), which illustrates individuals' behaviors toward an item, is calculated using SB values for each item by all individuals as follows.

$$RTF_i = \frac{\sum_{j=1}^{n} SB_{ij}}{N} \quad (2)$$

The value N in the formula represents the total number of individuals responding to the item. The Response Time Fidelity (RTF) index, developed based on individual behaviors, has been determined for each item, and the variation between years has been examined. As the items are common, the difference between RTF indices across cycles has been analyzed, and descriptive statistics, graphics, and estimates were created by using “data.table” package in the R software (Barrett et al., 2024). The mean response times of individuals and item characteristics in PISA 2015 and PISA 2018 were compared using a t-test. Spearman's correlation coefficient was used to estimate the correlation between variables. The Wilcoxon Test was used to compare response times of item characteristics over cycles. The information for each trend item has been compared with mean response times, and differences between cycles have been investigated.

Sample

In this study, data from trend mathematics items of Türkiye students used in the PISA 2015 and 2018 cycles were employed. The total number of Türkiye students who participated in the mathematics test was 5895 in 2015 and 6890 in 2018, including 50% female and 50% male students in 2015, and 49.6% female and 50.4% male students in 2018. Due to the presence of only 24 forms containing the 69 trend items related to mathematical literacy, and since not all students received the math items, data from 2408 students were used in PISA 2015 and data from 3718 students were used in PISA 2018. These math trend items were distributed in various forms due to matrix sampling. The sample size for each item ranged from 423 to 741 in 2015 and from 1100 to 1156 in 2018. Türkiye participated in the assessment in a computer-based format (MEB, 2019).

Results

This research examined the relationship between the characteristics of the 69 trend items in the mathematics subtest and item response times across the PISA 2015 and PISA 2018 cycles. Furthermore, the Response Time Fidelity (RTF) index for the trend items was investigated concerning the PISA 2015 and 2018 cycles.

The summary descriptive statistics of the mean response times given by individuals to the 69 trend items in the mathematics literacy subtest of PISA 2015 and PISA 2018 are presented in the table below.

Table 1.
The Summary Results of Item Response Times (in seconds) by year

	PISA 2015	PISA 2018
min	38.14	39.74
mean	99.38	118.26
median	95.03	112.39
max	179.33	208.28

The data in the table reveal an increase in students' response times to the items in 2018 compared to 2015. The mean response time expended by individuals across the 69 mathematics items was 99 seconds in 2015, while it reached 118 seconds in 2018. The mean response times of individuals in PISA 2015 and PISA 2018 were compared using a t-test. As a result of comparing the mean response times between the two groups, the estimated t-value was 3.49. This finding indicates a statistically significant difference in the mean response times between the two groups ($p < .05$).

The mean response time of items was compared across years based on item formats. The obtained results are presented in the table below.

Table 2.
Response Times According to Item Formats Across Years

		2015	2018		
Item Format	N	MRT (sec)	MRT (sec)	Z	p
Simple Multiple-Choice Computer Scored	16	82.20	98.48	-3.516	0.00
Complex Multiple-Choice Computer Scored	13	89.52	103.75	-3.296	0.00
Open Response Computer Scored	22	98.27	113.84	-4,107	0.00
Open Response Human Coded	18	123.13	151.73	-3,724	0.00

As seen in Table 2, the highest mean response times belong to open-ended items scored by scorers for both cycles. The lowest mean response time is for simple multiple-choice items automatically scored by the computer for both cycles. Across all item formats, response times were higher in 2018. Additionally, despite both being open-ended items, the response time for items scored by the computer is lower than the response time for items scored by scorers. To compare the response time difference according to item format, for all item formats, the Wilcoxon signed-rank test was conducted. Response time differences for all item formats over two PISA cycles are statistically significant.

The item response times related to mathematical processes, which are sub-dimensions of mathematical literacy assessment framework, were examined across years. The resulting outcomes are presented in the table below.

Table 3.
Response Times for Mathematical Process Across Years

		2015	2018		
Processes	N	MRT (sec)	MRT (sec)	Z	p
Formulating Situations Mathematically	21	95.62	116.88	-4.107	0.00
Interpreting, Applying and Evaluating Mathematical Outcomes	19	98.15	115.15	-3.823	0.00
Employing Mathematical Concepts, Facts and Procedures	29	102.91	121.30	-4.703	0.00

As seen in Table 3, the mean response time for items related to the employing mathematical concepts, facts, and procedures within mathematical processes is the highest in both cycles. The mean response time increased across all processes in the year 2018. To compare the response time difference across processes, the Wilcoxon signed-rank test was performed for all item processes. According to

mathematical processes of math items, the differences in response times are statistically significant between the two PISA cycles.

The relationship between item response times and item parameters was compared across years. The obtained results are shown in the table below. Item parameters including item difficulty and item discrimination index were calculated according to item response theory.

Table 4.
The Correlations Between Item Parameters and Item Response Times

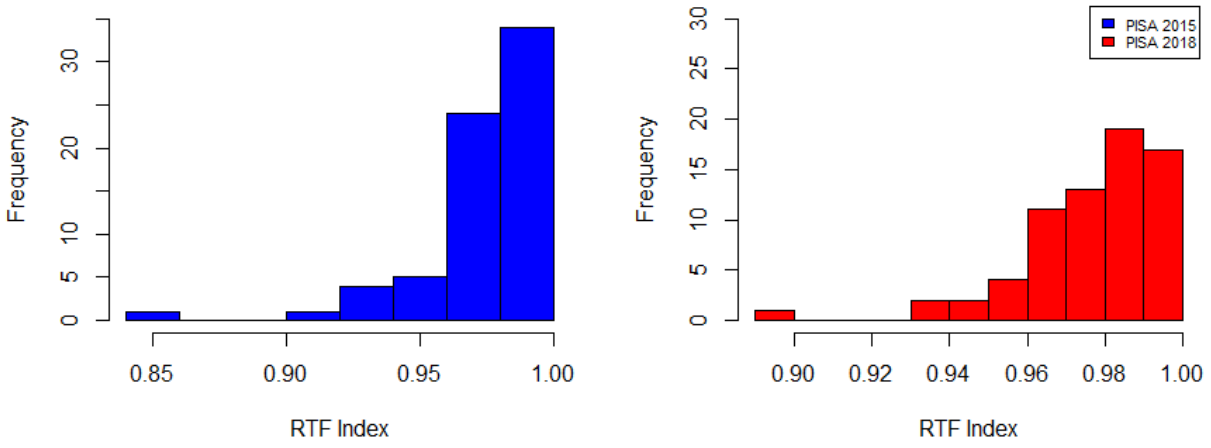
	Year	Item Difficulty	Item Discrimination
Item Response Time	2015	0,26**	0,20
	2018	0,27**	0,21

**p<0,01

Upon reviewing Table 4, a statistically significant positive correlation between the mean item response times and item difficulty is evident for both cycles. The correlation between the variables was moderate. It can be inferred that as items become more challenging, item response times tend to increase. However, no significant correlation was observed between item response times and item discrimination.

RTF indices for items were calculated based on the NT10 and NT20 methods for both cycles. According to the NT10 method, the mean RTF index for 2015 was 0.97. This suggests that 3% of students exhibited rapid-guessing behavior while answering the items. In 2018, the mean RTF index for items reached 0.98, with only 2% of students displaying rapid-guessing behavior. There was a decrease in the percentage of students exhibiting rapid-guessing behavior between the years. The histogram of item RTF indices based on cycles for the NT10 method is depicted below.

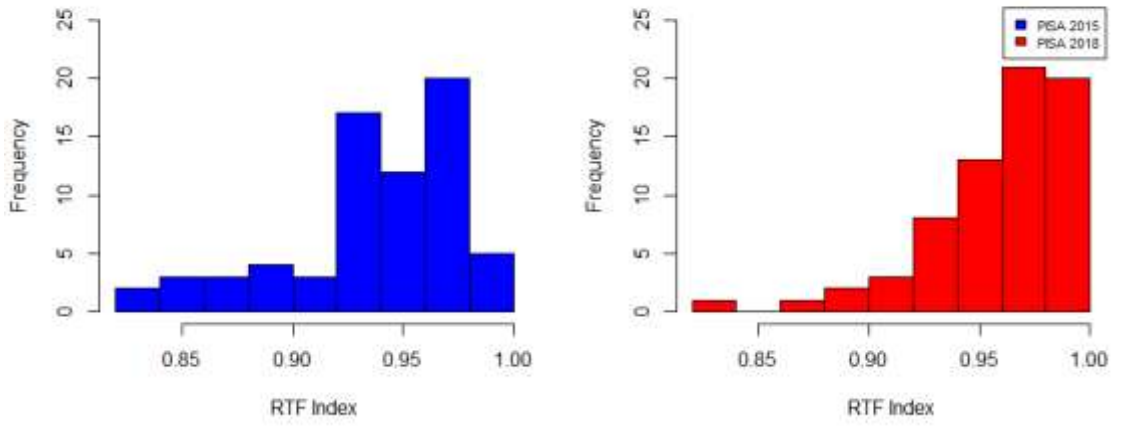
Figure 1.
Histograms of Item RTF Indices Based on Years According to the NT10 Method



According to Figure 1, the RTF indices are skewed to the left in both cycles. For 2015, the RTF indices range from 0.86 to 0.99, while for 2018, they vary between 0.89 and 0.99. Consequently, for each item, rapid-guessing behavior was exhibited by at least one individual.

Below is the histogram illustrating the item RTF indices based on cycles for the NT20 method.

Figure 1.
Histograms of Item RTF Indices Based on Years According to the NT10 Method



Based on the NT20 method, the mean RTF indices in 2015 was 0.94. Consequently, 6% of students showed rapid-guessing behavior while responding to the items. In 2018, the mean RTF index for items was 0.96, and the proportion of students displaying rapid-guessing behavior was 4%. There was a decline in the percentage of students exhibiting rapid-guessing behavior between the years according to this method.

According to Figure 2, the RTF indices obtained using the NT20 method are skewed to the left in both cycles. For 2015, the RTF indices range from 0.82 to 0.99, while for 2018, they vary between 0.84 and 0.99. The majority of individuals have exhibited solution behavior toward the items.

Discussion

In this study, the relationship between the characteristics of the 69 trend items in the mathematics subtest and item response times across the PISA 2015 and PISA 2018 cycles was examined. The mean response time for all items increased in the 2018 cycle, and this increase is statistically significant. Considering the relationship between response time and students' efforts toward the items, it can be suggested that in the 2018 administration, students exhibited more effort and solution behavior while answering the items. Türkiye's mathematics performance also improved in the PISA 2018 administration (MEB, 2019). Eklöf et al. (2014) found a statistically significant relationship between students' efforts obtained through self-reporting and their test performance scores. Therefore, the increase in Türkiye mathematics performance in PISA 2018 should be investigated to determine whether the increase is related to the increase in response time.

In the PISA, the mathematics item formats were analyzed according to the mean item response times, and the response times were compared between the two cycles. The item format that individuals spent the most time solving was open-ended items, while multiple-choice items required less time from individuals for both PISA cycles. Additionally, despite both being open-ended items, the response time for items scored by human raters was longer than for those scored automatically by computers for both PISA cycles. For all item formats, the mean item response time has increased, and response time differences in all item formats over two PISA cycles are statistically significant. A review of the literature revealed that the findings of this study align with the broader trends observed in previous research. Kuang and Sahin (2023) showed that disengagement has focused on one type of

disengagement, namely rapid-guessing, to multiple-choice items. In addition, Wise and Gao (2017) found that selected response item formats led to rapid-guessing and occasional rapid-omits. Yalçin (2022) showed that students spent more time on open-ended questions than multiple-choice questions. Birgili (2014) found that students exerted more effort when responding to open-ended questions compared to multiple-choice ones. Similarly, a study using TIMSS 2015 data (İlhan et al., 2020) revealed that students faced greater difficulty with open-ended items than with multiple-choice items.

Item Response Fidelity (RTF) indices, allowing determination of the proportion of behavior exhibited by individuals toward items, were established for each item. Additionally, RTF indices were compared between the two cycles. The RTF index indicating solution behavior exceeds 80% for each cycle. The majority of individuals displayed solution behavior towards the items. Similarly, the proportion of students displaying rapid-guessing behavior in the PISA 2018 administration decreased compared to the rate in the PISA 2015 administration.

The weighted sub-dimension within the mathematics assessment framework is centered around mathematical processes. These encompass "formulating situations mathematically", "employing mathematical concepts, facts, and procedures"' and "interpreting, applying, and evaluating mathematical outcomes" (OECD, 2023). Item response times were investigated within these processes. It was observed that items related to employing mathematical concepts, facts, and procedures had the longest mean item response times. Within the utilization process, the focus is on how individuals apply mathematical concepts, facts, and procedures in decision-making (MEB, 2019). Consequently, this phase demands individuals to engage their reasoning skills. Given that this skill involves problem analysis, correlating problem stages, making inferences, and proposing solutions, items related to the application of these reasoning skills may naturally require more time due to the complexity of such cognitive processes.

The relationship between mean item response times and item parameters was compared across the years. There was no significant relationship found between item response time and item discrimination in both cycles. However, there was a significant positive relationship between item response time and item difficulty for both cycles. Similar to these findings, Altuner (2019), demonstrated in their study a negative moderate relationship between item response time and item difficulty index and a significant relationship with the item discrimination index. Consistent with this study, item response time tends to increase for difficult items, while it significantly decreases for easier items.

The fact that PISA is not considered a high-stakes assessment for students may have influenced their response times and effort levels. Additionally, students in this study may not have been fully aware of the time constraints. It is important to note that the current study's findings are limited to students who answered mathematics trend items across two PISA cycles. Future research could extend this analysis to include items from all PISA tests to identify broader patterns. This study highlighted that various factors influence students' response times. Therefore, it is recommended that future studies examine the effect of response time on item and ability parameters. Additionally, leveraging recent technological advancements, such as eye-tracking devices, could provide a more detailed understanding of response times. Finally, this study used non-parametric tests due to the violation of normality assumptions and sample size. Should these assumptions be met in future studies, results could be compared using parametric methods.

In the study, a pattern of rapid-guessing behavior was observed more frequently in multiple-choice items, which indicates a high likelihood of guesswork. Therefore, it is recommended that the predominant use of short-response items, particularly for assessing the skills measured by the items where rapid-guessing is observed, be adopted. This approach could lead to more accurate estimations of students' ability levels.

Declarations

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as this study uses data shared with the public.

Author Contribution: Muhsin POLAT: conceptualization, investigation, methodology, data analysis, visualization, writing - review & editing. Hülya KELECİOĞLU: conceptualization, methodology, supervision, writing - review & editing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington DC.

Altuner, F. (2019). *Examining the relationship between item statistics and item response time* [Master's Thesis, Mersin University]. Retrieved from http://tez2.yok.gov.tr/

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., & Hocking, T. (2024). *data.table: Extension of 'data.frame'*. R package version 1.14.8. https://CRAN.R-project.org/package=data.table

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, 39(6), 502-523. doi:10.3102/1076998614558485

Debeer, D., Janssen, R., & Boeck, P. D. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, 54(3), 333-363. doi:10.1111/jedm.12147

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13(1), 55-77. doi:10.1207/s15324818ame1301_3

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross- national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, 27(1), 31-45. doi:https://doi.org/10.1080/08957347.2013.853070

Kuang, H., & Sahin, F. (2023). Comparison of disengagement levels and the impact of disengagement on item parameters between PISA 2015 and PISA 2018 in the United States. *Large-scale Assessments in Education*, 11(4). doi:10.1186/s40536-023-00152-0

Lee, Y. H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, 16(3), 240-267. doi:10.1080/15305058.2015.1085385

Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, 2(8), 1-24. doi:10.1186/s40536-014-0008-1

MEB. (2019). *PISA 2018 Türkiye Ön Raporu*. Ankara: Milli Eğitim Bakanlığı.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3-8. doi:10.1207/s15324818ame0901_2

Michaelides, M. P., & Militsa, I. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling*, 64(3), 304-338.

OECD. (2023). *OECD-PISA*. Retrieved from PISA 2018 Technical Report: https://www.oecd.org/pisa/data/pisa2018technicalreport/

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing*, 28, 31–56. doi:10.1007/s11145-014-9518-z

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, 9(10), 1-31. doi:10.1186/s40536-021-00104-6

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, 33(4), 263-279. doi:10.1080/08957347.2020.1789141

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34, 213-232. doi:10.1111/j.1745-3984.1997.tb00516.x

Setzer, J. C., Wise, S. L., Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, 1, 34-49. doi:10.1080/08957347.2013.739453

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384. doi:10.1007/s11336-007-9045-8

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, 19(2), 95-114. doi:10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19-38. doi:10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343-354. doi:10.1080/08957347.2017.1353992

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, 53(1), 86-105. doi:10.1111/jedm.12102

Wise, S. L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Paper presented at the 2012 annual meeting of the national council on measurement in education.* Vancouver: Canada.

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educational Measurement Issues and Practice*, 25(2), 21-30. doi:10.1111/j.1745-3992.2006.00054.x

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3);193-208

The Effect of Presenting Geometry Items with and without Shapes on the Psychometric Properties of the Test and Students' Test Scores*

İslim ATÇI ** Mustafa İLHAN ***

Abstract

This research aimed to examine the effects of presenting geometry items with and without shapes on the psychometric properties of the test and students' test scores. The study was conducted on 480 eighth grade students. Within the scope of the study, two geometry tests were crafted, one with shapes and the other without shapes. Both tests consisted of 15 multiple-choice items. In the data collection process, a counterbalanced design was followed, and the two tests were administered to the students three weeks apart. Analyses were carried out on 405 students who participated in both applications and whose test forms could be matched. The factor analysis results revealed that the factor loadings of the items and extracted variance were higher for the test with shapes compared to the test without shapes. The Cronbach's alpha coefficient of the test containing shapes was found to be significantly higher than that calculated for the test without shapes. According to item difficulties, the questions with shapes were easier for the students than the shape-free questions. In terms of discrimination indices, a difference in favor of the shape-containing test was observed in almost all items. Ferguson's delta statistic, which is a measure of discrimination for the overall test, was higher in the shape-containing test. Correlation analysis denoted a strong positive relationship between students' scores on the two tests. The paired samples *t*-test proved that there was a statistically significant difference between students' scores on the tests with and without shapes. These results indicate that the geometry tests with and without shapes differ in terms of both psychometric properties and students' test scores.

Keywords: Shape-containing geometry items, geometry items without shapes, psychometric properties

Introduction

Achievement is an abstract construct that cannot be directly observed but can be indirectly measured through tests. Therefore, reaching accurate estimations about individuals' achievement depends first and foremost on the quality of the test. There are two basic questions to be addressed when developing tests: (1) What are we going to measure, and (2) How can we measure this targeted characteristic (Lindquist, 1936)? In order to clarify what is to be measured, a test plan is usually prepared using a specification table. On the other hand, when it comes to the question of how to measure the targeted trait, various dilemmas arise (Rodriguez, 2002). These dilemmas may be related to item type, preparation of answer choices, or the structure of the item stem.

In item type dilemmas, the most appropriate item type (multiple-choice, true-false, open-ended, etc.) is decided by taking into account the construct being measured, the cognitive level of the learning objective being tested, and the number of examinees. In order to help the test developers/researchers in this decision process, many studies have been conducted to reveal how item type affects validity, reliability, item difficulty and discrimination, item response time, and examinees' ability scores (Bacon, 2003;

*This study has been produced from Master's Thesis that was conducted under the supervision of the Assoc. Prof. Dr. Mustafa İLHAN and prepared by İslim ATÇI.

** Teacher, 19 Mayıs Middle School, Ministry of National Education of the Republic of Türkiye, Mardin-Türkiye, islimatc@gmail.com, ORCID: 0009-0008-1729-4945
***Assoc. Prof. Dr., Dicle University, Ziya Gökalp Faculty of Education, Diyarbakır-Türkiye, mustafailhan21@gmail.com, ORCID: 0000-0003-1804-002X

To cite this article:
Atçı, İ. & İlhan, M. (2024). The effect of presenting geometry items with and without shapes on the psychometric properties of the test and students' test scores. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 193-208. https://doi.org/10.21031/epod.1483567

Received: 14.05.2024
Accepted: 24.09.2024

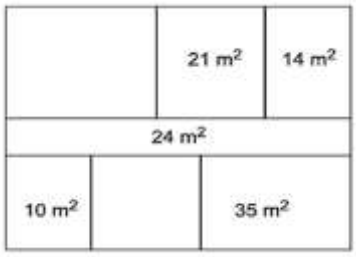
Cheng, 2004; Demir, 2010; Gültekin & Çıkrıkçı Demirtaşlı, 2012; İlhan et al., 2020; Yılmaz Koğar & Soysal, 2023; Öksüz & Güven Demir, 2019; Özer Özkan & Özaslan, 2018; Temizkan & Sallabaş, 2015; Zulaiha et al., 2021). Dilemmas about the answer choices focused on the effects of the following issues on measurements: optimal number of options (Atalmış, 2018; Baghaei & Amrahi, 2011; Haladyna & Downing, 1993; Nwadinigwe & Naibi, 2013; Raymond et al., 2019; Rodriguez, 2005; Vegada et al., 2016), the order in which options are presented (Cizek, 1994; Hohensinn & Baghaei, 2017; Karanfil & Neufeld, 2020; Lions et al, 2021; Lions et al., 2023; Shin et al., 2019), options' homogeneity (Ascalon et al., 2007; Atalmış & Kingston, 2018), and the options such as "all of the above" and "none of the above" (Atalmış & Kingston, 2017; Bishara & Lanzo, 2014; Crehan et al., 1993; Jonsdottir et al., 2021). When it comes to the item stem, in addition to issues that concern all disciplines such as the effects of item stem length (Abd El-Mohsen, 2008; Koepf, 2018), its completeness (in the form of a full or incomplete sentence) (Ascalon et al., 2007; Schaefer, 2009; Violato, 1991; Violato & Harasym, 1987; Violato & Marini, 1989) and orientation (negatively or positively) (Harasym et al., 1992; Harasym et al., 1993; Terranova, 1969) on psychometric qualities, field-specific dilemmas also arise. Mathematics is one of the disciplines where different dilemmas occur when writing item stem.

One of the basic dilemmas in item stem writing in mathematics tests is whether it would be more appropriate to compose the computational items with mathematical expressions or word problems, and whether such a change in the item stem would make a difference in the measurements (Kan et al., 2019). Another important dilemma appears in geometry, a sub-branch of mathematics. Just as items in the algebra and arithmetic areas of the mathematics tests can be written with word problems or mathematical expressions, geometry items can also be created with or without shapes. The two items in Figure 1, which the Ministry of National Education of Türkiye Republic included in the numerical ability test of the 2018 High School Entrance Examination, exemplify this.

Figure 1

The samples for geometry items with and without shapes

(a)



Above, the areas of some sections are given on the rectangular floor plan, where each section is rectangular.

If the side lengths of each of these rectangles are natural numbers in meters, the sum of the areas of the parts whose areas are not given is at least how many square meters?

A) 36 B) 54 C) 64 D) 76

(b)

A square-shaped garden with a side length of 10 m has an irrigation system only at the corners. Each irrigation system can irrigate up to a section up to 4 m away from its location. In the part of this garden that cannot be irrigated, there is a pergola with a square base. The diagonal of the base of this pergola coincides with the diagonal of the garden.

What is the maximum floor area of this pergola, whose base diagonal length is a natural number in meters?

A) 18 B) 48 C) 52 D) 72

As can be seen in the Figure 1.a, a geometric shape was presented in the item and the basic information related to the question was explained on this shape. In the question in the Figure 1.b, on the other hand, all information was given verbally and no geometric shape was provided. Such differences in the item stem can affect the cognitive processes that need to be employed to answer the item correctly (Kan et al., 2019). For example, the visuospatial skills needed to solve the geometry items with and without shapes may differ. In a similar vein, answering a geometry question that does not contain shapes and consists only of verbal expressions correctly may require more intensive verbal skills. In order to be able to develop more purposeful geometry tests and to read the measurement results more accurately, it is necessary to know the effect of such differences in the item stem on the measurements.

The Purpose and Importance of the Research

When writing geometry items, the test developer may encounter the following quandaries: *(a)* Should the shapes be presented compatible or incompatible with their actual values (Çetin & Türkan, 2013)?, *(b)* Should prototype drawings corresponding to the most familiar model of the geometric shape or non-prototypical drawings be employed? Certainly, the problem that is as important as these, perhaps even before these, is how the presentation of the item with shapes and only verbal expressions without shapes will affect the measurements. To put it more clearly, one of the research problems that needs to be answered is whether presenting geometry questions with or without shapes will make a difference in students' test scores and the psychometric properties of the test. However, when the literature is examined, it is seen that the number of studies on this subject is quite limited. One of these studies was conducted by Aydın et al. (2006) with 12th grade students, and students in the same class were randomly divided into two groups. Students in the first group answered the geometry test in which verbal expressions and shapes were presented together. The other group was administered a test consisting only of verbal expressions without shapes. As a result of the research, they determined that the averages in the group to which the shape-containing was applied were higher in all items. Karpuz et al. (2014), on the other hand, carried out a qualitative study to analyze students' responses to shape-containing and shape-free geometry questions comparatively. In the literature, no study was found to examine the impact of presenting geometry items with and without shapes on the psychometric properties of the test and whether it created a significant difference in students' test scores.

This empirical research attempts to determine the effect of presenting geometry items with or without shapes on the psychometric properties of the test and students' test scores. For this purpose, answers to the following problems were sought in the study.

1. Do the test forms in which geometry questions are presented with or without shapes differ in terms of *(a)* factor structures, *(b)* item difficulty and discrimination indices, and *(c)* internal consistency coefficients?
2. *(a)* What is the relationship between students' scores on geometry tests with and without shapes? *(b)* Is there a statistically significant difference between their scores of these two tests?

Since the past studies comparing the geometry tests with and without shapes have not dealt with these research problems, it is thought that this study has original value and will contribute to both mathematics education and measurement and evaluation literature. It is hoped that the study results will benefit teachers, mathematics education, and measurement and evaluation experts in the preparation of geometry tests and also indirectly shed light on the points to be considered in geometry teaching processes.

Methods

Research Design

The present study employed a descriptive-comparative design to contrast geometry tests with and without shapes. This research approach focuses on two variables, compares these variables by following a well-planned but not manipulated formal process, and aims to reveal which of the two variables/situations is better as a result of the comparison (Paler-Calmorin & Calmorin, 2007).

Participants

The study was carried out with eighth grade students because secondary school students were a more accessible group for the researcher who carried out the data collection process and because the number of objectives learned by eighth grade students in the field of geometry learning area was higher compared to secondary school students in lower grades. Accordingly, 480 eighth grade students from three different schools in Mardin province constituted the participants of the study. Nevertheless, 46 students who participated in one of the with and without shapes test administrations but did not

participate in the other, and 29 students whose forms could not be matched because they did not use the same nickname on the two tests they answered, were excluded from the analysis. Therefore, the analyses were carried out on a total of 405 students, 218 (53.83%) of whom were female and 187 (46.17%) of whom were male, who participated in both tests and whose answered test forms could be matched.

Instruments

Research data was collected via two tests prepared to cover the objectives of the 7th grade and the previous terms in the geometry learning field of the mathematics curriculum. Both tests had 15 multiple choice items. While the first test consisted of geometry questions with shape, in the second one the items were presented only with verbal expressions. The two test forms were parallel except that one was prepared with shapes and the other consisted of only verbal expressions without any shapes. In both tests, the items had four options. The order of the items and options, and the option corresponding to the correct answer were identical for the two tests.

After the draft form for the tests was created, opinions were received from two field experts. The first of the experts was a faculty member whose field of study includes geometry teaching, who gives courses on geometry teaching at the undergraduate level, and who has a doctorate in mathematics education. The second expert was a mathematics teacher with 10 years of professional experience. These two experts reviewed the items in terms of their suitability for the research purpose and scientific accuracy. Experts expressed their opinion that formal corrections were needed in the items. For example, they pointed out that there are differences in terms of font and font size from one item to another in the tests and that there should be a standard in this regard. Besides, they emphasized that the "x" symbol used to indicate angle or length was written with a capital letter in some questions and with a lowercase letter in some questions, and recommended the use of lowercase letters in all questions. Furthermore, the wording of some questions was changed based on the experts' opinions. Additionally, one of the field experts proposed that the first item may be removed from the tests, saying, "*This item will mostly be solved correctly, whether it is presented with or without a shape.*" However, considering that the tests were crafted according to the objectives of the 7th grade and previous years and that the students may have forgotten the topics, it was thought that the ease of the first item would increase students' motivation for the test. Hence, it was decided to keep the relevant item in the test.

Following the changes that were made based on the opinions of field experts, the opinion of an expert with the title of associate professor in the field of measurement and evaluation in education was consulted. This expert whose undergraduate was in the field of secondary school mathematics teaching reviewed both tests and he specified that the items were in accordance with measurement principles such as *(a)* emphasizing negative judgements, *(b)* listing the options from the largest to smallest or smallest to largest, and *(c)* ensuring a balanced distribution of the correct answer among the options. Finally, a Turkish teacher had a look at the tests in terms of spelling and punctuation rules, and necessary corrections were made to the items in line with her comments. Then, a preliminary trial was conducted on 12 students who were heterogeneous in terms of mathematics achievement. After answering the test, the students were interviewed, and there were no statements that the students had difficulty understanding in the items or in the test instructions. Thus, it was concluded that the tests were ready for application. An example of geometry items with and without shapes was presented in the Appendix.

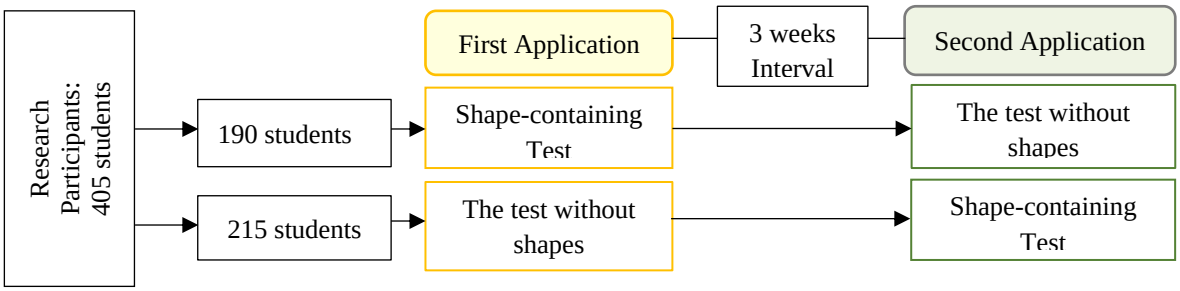
Data Collection Process

The data collection process was completed in two stages. In studies where two different instruments are administered to the same group at a certain interval, there may be effects arising from the order in which the instruments are applied, and this may threaten the internal validity of the research (Corriero, 2017). To prevent this, that is, to eliminate possible effects that may arise from the sequence of application of the tests, it is recommended to use a counterbalanced design (Graveter & Forzano, 2018). In this context, research data was collected according to the pattern summarized in Figure 2. Half of the group was first administered the shape-containing test and then the test without shapes, while the other half followed the reverse order.

Another important issue in studies involving repeated measures is the time elapsed between two applications. This period should be long enough that students do not remember their answers and short enough that participants do not experience changes due to maturation/learning (Crocker & Algina, 1986; Goldfarb, 2021). Since the time required to achieve this will vary depending on the developmental characteristics of the participant group and the nature of the measured construct (Mitchell et al., 2000), there is no clear opinion on how much time should be between two applications. However, a period of 2-3 weeks is generally considered ideal for achievement tests. Therefore, in the current study, two test forms were administered to the students three weeks apart.

Figure 2

The path followed in the collection of the research data



Before the data collection process, approval was received from Dicle University, Social and Human Sciences Ethics Committee regarding the compliance of the study with current scientific ethical principles. In the next step, a preliminary application was made for Research, Competition and Social Event permission on the Ministry of National Education website. After the application, the leave petition was submitted for approval by the Mardin/Artuklu District Governorship. Once all necessary permissions obtained, data collection started in secondary schools in Artuklu district of Mardin province, and the applications were carried out in the classroom environment, in paper-pencil form and on a voluntary basis, between December 2023 and January 2024. Prior to application, students were informed about the aim of the study and it was emphasized that the data would be used only for scientific purposes and would not be shared with any other person or institution. In addition, students were reminded that they did not need to write their actual names on the test forms, but it was stated that they should write a nickname that they would not forget in the space provided at the beginning of the tests in order to match the two test forms they would answer. There was no student who refused to participate in the study in any classroom where the application was carried out.

Data Analyses

Procedures for data analysis were presented under three headings: preliminary analyses, validity and reliability analyses, and analysis for comparing the students' scores in two tests.

Preliminary Analyses

This title includes the processes carried out to prepare the data sets for analysis and the results of the analyses applied to check the distribution of the data. While performing the analyses in question, the JASP 0.18.1.0 program (JASP Team, 2022) and the web tool running R software in its background developed by Aybek (2021) were utilized. Since the multiple-choice tests were employed as the instruments in the study, correct answers were scored as 1, and incorrect answers and blank items were scored as 0. Therefore, there were no missing values in the data file. Moreover, univariate and multivariate outliers were not found in the data set. After this determination, skewness and kurtosis coefficients were examined for univariate normality and Henze–Zirkler statistics for multivariate normality. Table 1 depicts the results of the normality test.

Table 1

The results for univariate and multivariate normality tests

Test	Skewness		Kurtosis		Henze–Zirkler
	Statistic	Std. Error	Statistic	Std. Error	
With shapes	.14	.12	-1.15	.24	4.45*
Without shapes	.76	.12	.02	.24	2.22*

*p<.01

The fact that the skewness and kurtosis coefficients are within ±1.5 is judged as the evidence of univariate normality (Tabachnick & Fidell, 2013). Accordingly, it is understood that the research data meet the assumption of univariate normality. The statistical significance of the Henze Zirkler test, on the other hand, indicates that multivariate normality is violated.

Validity and Reliability Analyses

In the research, Exploratory Factor Analysis (EFA) was performed to ascertain the factor structure of the tests. In EFA, the Kaiser–Meyer–Olkin (KMO) values were found to be .781 and .541 for the shape-containing test and shape-free test, respectively. Besides, Bartlett's sphericity test results were significant for both forms [$\chi^2_{shape-containing\ test}$=3634.221, $\chi^2_{shape-free\ test}$= 2386.016; *df* = 105; p < .001]. The calculated KMO values over .50 and the statistically significant Bartlett's tests reflect that the sample is adequate and the correlation matrices are suitable for acquiring reliable factors (Field, 2013). Thereby, the analysis continued and since the multivariate normality assumption was violated, the principal axis factoring technique (Şahin, 2022), which does not require any prerequisites about the distribution of the data, was operated as the estimator in EFA. Parallel analysis method was used to decide the number of factors, and analyses were carried out based on the tetrachoric correlation matrix as the data had a dichotomous (1-0) structure.

Subsequently, the two test forms were compared in terms of item difficulty and discrimination indices, and reliability. For the items' discrimination, the discrimination index (r_{jx}) based on 27% lower-upper group comparisons and the point biserial correlations (r_{pb}) were calculated. Also, in order to attain a discrimination index for the entire test, Ferguson's delta (δ) statistic (Ferguson, 1949) was calculated using Equation 1, where *k* is the number of items, *n* is the number of test takers (i.e., sample size) and *f* is the frequency value of each score. Ferguson's delta provides information about how heterogeneous the examinees' test scores are (Zhang & Lidbury, 2013) and can take values ranging from 0 to 1 (Hernandez and Zalava, 2017). The value of .90 is recommended as the threshold for this statistic (Kline, 1993).

$$\delta = \frac{(k+1)(n^2 - \sum f^2)}{kn^2} \qquad (1)$$

Within the scope of reliability analysis, Cronbach's alpha internal consistency coefficients were calculated and the significance of the difference between the Cronbach's alpha values of the two forms was tested using the method recommended by Feldt et al. (1987). In psychometric analyses, JASP 0.18.1.0 program (JASP Team, 2022) was utilized for EFA. While reliability coefficients and item statistics were calculated in TAP (Test Analysis Program) software (Brooks & Johanson, 2003), Ferguson's delta statistics were computed in Microsoft Excel. To compare Cronbach's alpha coefficients statistically, on the other hand, the interface running the cocron package in R programming language, developed by Diedenhofen and Musch (2016) was used.

Analyses to Compare Students' Scores in Two Tests

Since the research data held the univariate normality assumption, the relationship between student scores in the shape-containing and shape-free tests was examined by means of Pearson product-moment correlation. When interpreting the size of correlation coefficient, the following ranges offered by Salkind (2010) were taken as reference: between .00 and .20, very weak; between .20 and .40, weak, between .40 and .60, moderate; between .60 and .80, strong; between .80 and 1.00, very strong. Paired samples *t*-test was implemented to test the significance of the difference between the scores of the students from the two tests. In order to evaluate the magnitude of the significant difference observed as a result of the *t*-test, Cohen's *d* statistic was inspected. Cohen (1977) defined the cut-off points for small, medium, and large effects as .20, .50, and .80, respectively. Relying on this guideline, Cohen's *d* was interpreted as follows in the current research: if $d < .20$ the difference is negligible, if $.20 < d < .50$ the difference is small, if $.50 < d < .80$ the difference is moderate and if $d > .80$ the difference is large. Analyses to compare students' scores in the two tests were conducted in the JAPS 0.18.1.0 program.

Results

Firstly, EFA was applied for the tests with and without shapes. Table 2 shows the outputs reported in parallel analysis for the number of factors in EFA.

Table 2
Results from parallel analysis for number of factors in the tests with and without shapes

	Test with shapes		Test without shapes	
	Eigenvalues for Real Data	Eigenvalues for Simulated Data	Eigenvalues for Real Data	Eigenvalues for Simulated Data
Factor1	7.076*	1.348	4.720*	1.348
Factor2	1.201	1.264	1.672*	1.264
Factor3	0.967	1.204	1.418*	1.204
Factor4	0.873	1.158	1.208*	1.158
Factor5	0.853	1.113	0.973	1.113

Table 2 illustrates that the number of factors where the eigenvalue of the real data is greater than the eigenvalue of the simulated data was 1 in the test with shapes and 4 in the test without shapes. In other words, there was a unidimensional structure in the shape-containing test and a four-factor structure in the test without shapes. However, no interpretable structure was observed when the distribution of the items to the factors in the shape-free test was examined. More explicitly, the four factors that emerged could neither be associated with the theoretical framework such as the objectives measured by the items, nor with statistical features such as the items' difficulty indices. For this reason, considering that all items were written in a way to test the objectives belonging to the same learning domain, the number of factors was limited to 1 for the shape-free test and EFA was redone. Table 3 displays the factor analysis results obtained for the tests with and without shapes after the repeated EFA.

As can be seen from Table 3, the factor loadings of all items except Item 8 in the test without shapes are above the threshold value of .30 (Büyüköztürk, 2010). It is also noteworthy that the factor loadings of the items are generally higher in the shape-containing test compared to the test without shapes. In line with the factor loadings, the extracted variance ratio was also higher in the shape-containing test than in the shape-free one.

Table 3

Factor solutions reported in EFA for tests with and without shapes

Items	Factor Loadings	
	Test with shapes	Test without shapes
Item1	.676	.397
Item2	.595	.474
Item3	.527	.589
Item4	.495	.549
Item5	.767	.494
Item6	.624	.432
Item7	.499	.382
Item8	.781	.262
Item9	.599	.511
Item10	.715	.605
Item11	.686	.687
Item12	.730	.592
Item13	.796	.566
Item14	.595	.358
Item15	.715	.681
Extracted Variance	43.60%	26.90%

Following the factor analysis, item difficulty and discrimination indices were examined. Table 4 provides the results regarding item statistics.

Table 4

The results of the item analysis for the tests with and without shapes

Items	Test with shapes			Test without shapes		
	p	r_{jx}	r_{pb}	p	r_{jx}	r_{pb}
Item1	.85	.37	.48	.85	.26	.30
Item2	.54	.58	.53	.34	.46	.45
Item3	.33	.52	.47	.38	.48	.51
Item4	.64	.56	.46	.43	.53	.49
Item5	.53	.78	.66	.32	.46	.48
Item6	.48	.67	.56	.51	.52	.43
Item7	.64	.53	.46	.54	.45	.39
Item8	.66	.74	.64	.32	.23	.31
Item9	.64	.63	.53	.46	.61	.50
Item10	.59	.73	.62	.57	.64	.53
Item11	.37	.69	.58	.30	.58	.56
Item12	.47	.79	.63	.33	.52	.50
Item13	.51	.81	.67	.28	.54	.49
Item14	.44	.69	.54	.28	.37	.38
Item15	.51	.75	.62	.38	.66	.56
Mean	.547	.563	.655	.413	.490	.458

It can be seen from the statistics in Table 4 that compared to the test without shapes, the difficulty indices in the shape-containing test are closer to 1 in most items. Accordingly, it is understood that among the geometry questions that aim to test the same learning objectives, the ones without shapes are more difficult for the students than the ones with shapes. This can be seen more clearly in the mean difficulties calculated for the tests. Table 4 shows that the discrimination indices based on upper-lower group comparisons exceeded the .30 lower limit (Ebel & Frisbie, 1991; Erkuş, 2012) in all items except Item 1 and Item 8 in the test without shapes. The discrimination values based on point biserial correlation

meet the .30 criterion for all items in both test forms. These results reflect that both shape-containing and shape-free tests have acceptable discrimination. However, both the item-based discrimination indices and the mean discrimination values of the tests indicate that the form with shapes can distinguish students at different achievement levels better than the shape-free one.

In the present study, Ferguson's delta (δ) statistic was also explored to obtain additional evidence about the difference between the discrimination powers of the tests and it was found .986 and .961 for the tests with and without shapes, respectively. That's to say, Ferguson's delta statistic exceeds the cut-off value of .90 for both tests. These results hint that the scores of both shape-containing and shape-free tests are heterogeneous enough to assert that the instruments are distinctive. Nevertheless, the greater Ferguson's delta statistic regarding the test with shapes implies that this test is more discriminatory than the test without shapes. Following the analysis of the distinctiveness of the tests, the internal consistency of the measurements obtained from the two tests was scrutinized. Table 5 shows the Cronbach's alpha coefficients calculated for the tests with and without shapes, along with the chi-square value for the significance of the difference between these coefficients.

Table 5

Internal consistency coefficients calculated for the tests with and without shapes

Tests	Cronbach's alpha	n	df	χ^2
With Shapes	.847 (95% CI [.825, .867])	405	1	12.459*
Without Shapes	.734 (95% CI [.696, .770])			

*$p < .001$

Table 5 denotes that Cronbach's alpha coefficients are above .70 (Pallant, 2005), which is the most commonly accepted lower limit for reliability for both tests. The Cronbach's alpha value for the test with shapes was higher than the test without shapes, and the difference between the internal consistency coefficients was statistically significant. Finally, the relationship between the scores the students received from the two tests was examined and whether the difference between the scores was significant or not was explored. Table 6 contains the results of the correlation analysis and paired samples *t*-test applied for this purpose.

Table 6

The results of correlation analysis and paired samples t-test for the scores of the tests with and without shapes

Tests	Mean	SD	Pearson r	Paired samples *t*-test		
				t	df	Cohen d
With Shapes	8.202	4.079	.648* (%95 CI [.588, .701])	12.78*	404	.635
Without Shapes	6.195	3.236				

* $p < .001$

As seen in Table 6, students' scores in the test with shapes were higher than the test without shapes. The calculated correlation coefficient reflects that there was a strong positive relationship between the students' scores in the two tests. The outputs of paired samples *t*-test indicated that there was a statistically significant difference between students' scores in the tests with and without shapes. The Cohen's *d* statistic represents a medium-sized difference, thus signs that the statistically significant difference detected was also noteworthy in practice.

Conclusion and Discussion

In this study, the effects of presenting geometry items with and without shapes on the psychometric properties of the test and students' test scores were examined. First, two tests were compared in terms of factor structures. The results showed that the factor loadings and extracted variance of the shape-containing test were higher compared to the form without shapes. This result reflects that the test with shapes serves the purpose more, in other words, it produces more valid measurements. Since there is no

visual support for the students in the test without shapes, variables such as language skills and reading comprehension ability may have a greater impact on the students' performance in this test compared to the shape-containing one. The interference of such sources of variability in the measurement results other than geometry knowledge may have decreased the validity of the measurements. The fact that a multidimensional structure emerged in the shape-free test when there was no limitation on the number of factors supports this view. As a matter of fact, Kan et al. (2019) compared the tests in which the item stem was presented in mathematical expressions and verbal form in terms of dimensionality and found that the two item types differed in terms of the skills required to answer the question correctly.

When the two forms were compared in terms of item difficulty indices, it was disclosed that the test with shapes was easier for the students than the one without shapes. That is to say, while students showed higher success in the form with shapes, they had difficulty in solving the questions when the same items were presented with only verbal expressions. This finding is coherent with the results of the study conducted by Karpuz et al. (2014). Karpuz et al. (2014) prepared two tests, one in which the concept and the shape were presented together, and the second in which the concept was presented but the shape was not. They administered these tests, both of which contained eight open-ended questions, to 120 high school students, one month apart, first the form without shapes and then the form with shapes. As a result of the study, they reported that students were more successful in solving the shape-containing questions. Drawing wrong shapes that do not meet the generalizability condition, being mostly influenced by prototype shapes while solving, or not being able to create any shape that corresponds to the conceptual knowledge in the item were listed as the factors that caused students to have more difficulty in solving the questions without shapes. A similar result was obtained in the study by Aydın et al. (2006). In the study just mentioned, an application was made over five open-ended geometry questions and students were randomly divided into two groups in the classroom environment. The students in the first group were presented the items with both verbal expressions and shapes. The students in the second group were asked the same questions without shapes. As a result of the study, they observed that when the item was presented only with verbal expressions, students had difficulty in transferring the expression in the question to the shape and consequently had more difficulty in these types of questions. The attained results regarding the item difficulties also match with the positions of Michael–Chrysanthou et al. (2024) who stated that "*A figure is a representation of a geometrical situation easier to understand compared to a representation with linguistic elements only*." Therefore, it can be said that the item difficulties calculated in this study for the tests with and without shapes are in line with the results of the previous studies.

The fact that students have more difficulty with the geometry questions without shapes can also be explained by the fact that they rarely encounter these types of questions. Because when exams do not go beyond certain question types, namely, when students always see similar types of items, they may have difficulty with different types of questions (Yılmaz, 2007). This situation manifested itself in the opinions written by the students under the questions in the test without shapes. Although there was not a particular space on the tests for students to write their comments about the questions, some students expressed that they did not know what to do with the shape-free questions and wrote notes on their test papers such as "*I can't understand as there is no shape*", "*algebraic expressions were already all we need in the geometry questions*", "*what kind of geometry question are these"*. Based on these opinions, it can be asserted that students' unfamiliarity with shape-free geometry items caused them to have more difficulty with these questions.

Another noteworthy result regarding item difficulties was as follows: The difficulty indices calculated for the first item in the tests with and without shapes were equal to each other. Indeed, one of the field experts whose opinion was consulted about the tests remarked that this item would be solved correctly whether it was presented with or without a shape. In this sense, the difficulty index calculated for the related item confirmed the expert opinion. Accordingly, it may be useful to get the experts' opinions before the application about the necessity of the shape in geometry questions or in which questions removing the shape may make a difference in the difficulty index.

When the item discriminations calculated for the two tests were compared, a difference was observed in favor of the shape-containing form. This result reflects that when geometry questions are presented with shapes, students with different levels of achievement can be effectively discriminated from each other, whereas when the same questions are presented only with verbal expressions, it becomes difficult to distinguish students at different achievement levels. In line with this, the Ferguson's delta statistic, which provides information about how heterogeneous the examinees are in terms of their test scores, was higher for the shape-containing test. This difference between the discrimination powers of the tests was reflected in Cronbach's alpha coefficients, and a significantly higher internal consistency coefficient was estimated for the shape-containing test compared to the one without shapes. Accordingly, it is possible to conclude that there is a higher consistency between the items of the shape-containing test and that the shape-free form is more prone to random errors than the one with shapes. In the literature, there is no study directly comparing geometry tests with and without shapes in terms of discrimination and internal consistency. However, considering that there is a difference between the discrimination values and internal consistency coefficients of the tests even when geometric shapes are presented in accordance with their real values and different from their real values (Çetin & Türkan, 2013), it is thought that presenting the questions without shapes will affect these statistics more explicitly. Therefore, it can be said that the results of the study conducted by Çetin and Türkan (2013) indirectly support the findings of the study, although not directly.

In the second problem of the study, students' scores in tests with and without shapes were compared. The obtained correlation coefficient elicited that there was a strong positive relationship between the students' scores in the two tests. This result indicates that the two tests ranked the students largely similarly in terms of their geometry achievement. More clearly, there was a high relative agreement between the achievement scores obtained from the geometry tests with and without shapes. On the other hand, there was a statistically significant difference between the students' scores in the two test forms. This finding signs that there is no absolute agreement between the scores of the two tests. Likewise, Aydın et al. (2006) reported that the students' item scores were higher in the test with shapes compared to the one without shapes.

The fact that the students' scores in the test without shapes were significantly lower means that they could not use their conceptual knowledge in questions without shapes, had difficulty in creating visual representations of verbal expressions and were unable to mobilize the knowledge in their minds when they encountered questions without shapes. As a matter of fact, Çiftçi and İşleyen (2022) stated that students comprehend geometry problems in a shape-oriented manner, cannot transfer the verbal expressions to the shapes or they transfer them incorrectly, and that some students even skip the questions presented only verbally without reading them at all. In addition, they emphasized that the fact that students proceed directly through shapes without fully learning geometric concepts comes to light by the deficiencies in visualizing verbal expressions. In a similar vein, Barut and Retnawati (2020) showed the lack of visualization ability as one of the difficulties experienced in geometry lessons in their study conducted with secondary school students. Considering all these, it can be argued that one of the main factors that led students to get lower scores in the shape-free test is the deficiency in visualization skills, which Duval (1998) defines as one of the three basic cognitive processes of geometry teaching (cited in Çiftçi & İşleyen, 2022).

Implications, and Suggestions for Future Researches

The research results demonstrated that geometry tests, in which items are presented with or without shapes, differ in terms of both their psychometric properties and the test scores of the students. From the point of these results, it is possible to offer the following suggestions for practice: First and foremost, when experts from the field of mathematics education and measurement and evaluation need to prepare parallel forms of a geometry test, they should take into account that shape-containing and shape-free questions to test the same learning objective are not equivalent. Considering that students' performance on geometry items without shape is lower, teachers should focus more on conceptual learning in the lesson and provide opportunities for students to draw the shape of a geometric term given a definition.

Supporting the interaction between concept and shape with activities can improve students' visualization and spatial thinking skills and promote their performance on geometry questions without shapes. A similar situation is also valid for the textbooks. Including shape-containing geometry questions as well as shape-free items in textbooks may support students' visualization skills and prevent them from floundering when they encounter shape-free questions.

While interpreting the study results and implications based on these results, it should be kept in mind that the research has certain limitations and further research is needed to overcome these limitations. First of all, the current study was limited to the data obtained from two 15-item tests, one consisting of shape-containing and the other consisting of shape-free questions for eighth grade level. Therefore, it may be recommended to conduct a similar study with students at different grade levels. In addition, in the present investigation, no opinion was requested from the field experts about whether the shape would be necessary or not in the items prepared. In future studies, experts can be consulted, and it can be tested whether the differences found between the item statistics of shape-containing and shape-free questions are compatible with the experts' opinions about the necessity of shape. Finally, it can be tested whether the difference between the item statistics of questions with and without shapes changes according to whether the shape in the question is prototypical or unusual.

Declarations

Author Contribution: İslim ATÇI: conceptualization, methodology, development of the instruments, data collection and analysis, writing, and visualization. Mustafa İLHAN: determining the research problem, methodology, data analysis, writing-review & editing, and supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Ethical rules were followed in this research. Ethical approval for the study was received from Dicle University, Social and Human Sciences Ethics Committee dated 10.11.2023 numbered 598100.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Abd El-Mohsen, M. M. (2008). *The effect of stem length in multiple choice questions on item difficulty in syllabus-based vocabulary test items difficulty in syllabus-based vocabulary test items* [Unpublished Master Theses, The American University in Cairo]. Retrieved from https://fount.aucegypt.edu/retro_etds/2195/

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*(2), 153–170. https://doi.org/10.1080/08957340701301272

Atalmış, E. H. (2018). The use of three-option multiple choice items for classroom assessment. *International Journal of Assessment Tools in Education*, *5*(2), 314–324. https://doi.org/10.21449/ijate.421167

Atalmış, E. H., & Kingston, N. (2017). Three, four, and none of the above options in multiple-choice items. *Turkish Journal of Education*, *6*(4), 143–157. https://doi.org/10.19128/turje.333687

Atalmış, E. H., & Kingston, N. M. (2018). The impact of homogeneity of answer choices on item difficulty and discrimination. *Sage Open*, *8*(1). https://doi.org/10.1177/2158244018758147

Aybek, E. C. (2021). *Data preparation for factor analysis*. https://shiny.eptlab.com/dp2fa/

Aydın, E., Kertil, M., Yılmaz, K., & Topçu, T. (2006). *Examining contextual support in geometry learning in terms of student and question level* [*Geometri öğreniminde bağlamsal desteğin öğrenci ve soru seviyesi açısından incelenmesi]* [Full text oral presentation]. VII. National Science and Mathematics Education Congress, Gazi University, Gazi Education Faculty, Ankara.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31–36. https://doi.org/10.1177/0273475302250570

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*(2), 192–211.

Barut, M. E. O., & Retnawati, H. (2020). Geometry learning in vocational high school: Investigating the students' difficulties and levels of thinking. *Journal of Physics: Conference Series 1613*(1), 012058. https://doi.org/10.1088/17426596/1613/1/012058

Bishara, A. J., & Lanzo, L. A. (2014). All of the above: When multiple correct response options enhance the testing effect. *Memory*, *23*(7), 1013–1028. https://doi.org/10.1080/09658211.2014.946425

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

Büyüköztürk, Ş. (2010). *A manual of data analysis for social sciences* [*Sosyal bilimler için veri analizi el kitabı]* (11. ed). Pegem Academy.

Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*(4), 544–553. https://doi.org/10.1111/j.1944-9720.2004.tb02421.x

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*(1), 8–20. https://doi.org/10.1177/0013164494054001002

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic.

Corriero, E. (2017). Counterbalancing. In *The SAGE Encyclopedia of Communication Research Methods* (Vol. 4, pp. 278–281). Sage. https://doi.org/10.4135/9781483381411

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*(1), 241–247. https://doi.org/10.1177/0013164493053001027

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. HarcourtBrace Jovanovich.

Çetin, B., & Türkan A., (2013). The effect of the compatibility and incompatibility of the shapes with their actual values in secondary school 8th grade geometry test questions on the psychometric properties of the test [İlköğretim 8. sınıf geometri testi sorularında şekillerin gerçek değerlerine uygun çizilmesiyle, farklı çizilmesinin testin psikometrik özelliklerine etkisi]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*. *4*(2), 52–63. https://doi.org/10.21031/epod.77190

Çiftçi, O., & İşleyen, T. (2022). Üçgenin açıortayları ve kenarortayları konusunda öğrencilerin karşılaştıkları öğrenme güçlükleri. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, *23*(Özel Sayı), 509–560. https://doi.org/10.29299/kefad.943663

Demir, E. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) bilişsel alan testlerinde yer alan soru tiplerine göre Türkiye'de öğrenci* (Tez No. 257803), [Yüksek lisans tezi, Hacettepe Üniversitesi]. YÖK Ulusal Tez Merkezi.

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

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

Erkuş, A. (2012). *Measurement and scale development in psychology–I: Basic concepts and procedures [Psikolojide ölçme ve ölçek geliştirme-1: Temel kavramlar ve işlemler]*. Pegem Academy.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Ferguson, G. A. (1949). On the theory of test discrimination. *Psychometrika*, *14*(1), 61–68. https://doi.org/10.1007/bf02290141

Field, A. (2013). Discovering statistics using SPSS (3rd ed.). Sage.

Goldfarb, R. (2021). *Consuming and producing research in communication sciences and disorders: Developing power of professor*. Plural.

Graveter, F. J., & Forzano, L. B. (2018). *Research methods for the behavioral sciences* (6th ed.). Cengage.

Gültekin, S., & Demirtaşlı, N. Ç. (2012). Comparing the test information obtained through multiple choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online*, *11*(1), 251–263.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999–1010. https://doi.org/10.1177/0013164493053004013

Harasym, P. H., Doran, M. L., & Brant, R., & Lorscheider, F.L. (1993). Negation in stems of single-response multiple-choice items: An overestimation of student ability. *Evaluation & the Health Professions*, *16*(3), 342–357. https://doi.org/10.1177/016327879301600307

Harasym, P. H., Price, P. G., Brant, R., Violato, C., Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions, 15*(2), 198–220. https://doi.org/10.1177/016327879201500205

Hernandez, E. & Zalava, G. (2017). Accurate items for inaccurate in undergraduate physics students. In M.S. Ramírez- Montoya (Eds.), *Handbook of research on driving STEM learning with educational technologies* (pp. 315-340). IGI Global.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, *38*(1), 93–109.

İlhan, M., Boztunç Öztürk, N., & Şahin, M. G. (2020). The effect of the item's type and cognitive level on its difficulty index: The sample of TIMSS 2015. *Participatory Educational Research, 7*(2), 47–59. https://doi.org/10.17275/per.20.19.7.2

JASP Team (2022). *JASP (Version 0.18.1.0)* [Computer software]. https://jasp-stats.org/

Jonsdottir, A. H., Jonmundsson, T., Armann, I. H., Gunnarsdottir, B. B., & Stefansson, G. (2021, 8-9 March). *The effect of the number of distractors and the "none of the above" – "all of the above" options in multiple choice questions* [Conference presentation]. 5th International Technology, Education and Development Conference. https://doi.org/10.21125/inted.2021.1540

Kan, A., Bulut, O., & Cormier, D. C. (2019). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*, *24*(1), 13–32. https://doi.org/10.1080/10627197.2018.1545569

Karanfil, T., & Neufeld, S. (2020). The role of order and sequence of options in multiple-choice questions for high-stakes tests of English language proficiency. *International Journal of Applied Linguistics and English Literature*, *9*(6), 110–129. https://doi.org/10.7575/aiac.ijalel.v.9n.6p.110

Karpuz, Y., Koparan, T., & Güven, B. (2014). Students' use of shape and concept knowledge in geometry [Geometride öğrencilerin şekil ve kavram bilgisi]. *Turkish Journal of Computer and Mathematics Education*, *5*(2), 108–118. https://dergipark.org.tr/en/pub/turkbilmat/issue/21573/231505

Kline, P. (1993). *Handbook of psychological testing* (2nd ed.). Routledge.

Koepf, T. M. (2018). *The effect of item stem and response option length on the item analysis outcomes of a career and technical education multiple choice assessment* [Unpublished Doctoral Dissertation, Western Michigan University]. Retrieved from https://scholarworks.wmich.edu/dissertations/3366/

Lindquist, E. F. (1936). The theory of test construction. In H. W. Hawkes, E. F. Linquist & C. R. Mann (Eds.), *The construction and use of achievement examinations: A manual for secondary school teachers* (pp. 17–106). Houghton Mifflin.

Lions, S., Dartnell, P., Toledo, G., Godoy, M. I., Córdova, N., Jiménez, D., & Lemarié, J. (2023). Position of correct option and distractors impacts responses to multiple-choice items: Evidence from a national test. *Educational and Psychological Measurement*, *83*(5), 861–884. https://doi.org/10.1177/00131644221132335

Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., Córdova, N., Jiménez, D., Blanco, M. P., Ortega, G., & Lemarié, J. (2021). The position of distractors in multiple-choice test items: The strongest precede the weakest. *Frontiers in Educiton*, *6*, 731763. https://doi.org/10.3389/feduc.2021.731763

Michael–Chrysanthou, P., Panaoura, A., & Gagatsis, A. (2024). Exploring secondary school students' geometrical figure apprehension: cognitive structure and levels of geometrical ability. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-024-10317-5

Ministry of National Education of Türkiye Republic. (2018). *Central examination for secondary education institutions that will accept students by test: Numerical part.* Retrieved from https://odsgm.meb.gov.tr/meb_iys_dosyalar/2018_06/03153730_SAYISAL_BYLYM_A_kitapYY.pdf

Mitchell, J. E., Crosby, R. D., Wonderlich, S., & Adson, D. E. (2000). *Elements of clinical research in psychiatry*. American Psychiatric.

Nwadinigwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test ıtem and the psychometric characteristics. *Journal of Education and Practice*, *4*(28), 189–196. Retrieved from https://www.iiste.org/Journals/index.php/JEP/article/view/9944

Öksüz, Y., & Güven Demir, E. (2019). Comparison of open ended questions and multiple choice tests in terms of psychometric features and student performance. *Hacettepe University Journal of Education*, *34*(1), 259–282. https://doi.org/10.16986/HUJE.2018040550

Özer Özkan, Y., & Özaslan, N. (2018). Student achievement in Turkey, according to question types used in PISA 2003-2012 mathematic literacy tests. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(1), 57–64. https://pdfs.semanticscholar.org/7e84/37899e70c78fbe2dde7ab179ccca7eb6a0a0.pdf

Paler-Calmorin, L., & Calmorin, M. A. (2007). *Research methods and thesis writing* (2nd ed.). Rex Book Store.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows (Version 12).* Allen & Unwin.

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Science Education*, *24*, 141–150. https://doi.org/10.1007/s10459-018-9855-9

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Lawrence Erlbaum Associates.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Salkind, N. J. (2010). *Statistics for people who (think they) hate statistics* (3rd ed.). Sage.

Schaefer, J. M. L. (2009). *The effects of stem completeness and stem orientation on multiple-choice item difficulty and discrimination* [Unpublished Master Theses, California State University]. Retrieved from https://hdl.handle.net/10211.9/162

Shin, J., Bulut, O., & Gierl, M. J. (2019). The effect of the most-attractive-distractor location on multiple-choice item difficulty. *The Journal of Experimental Education*, *88*(4), 643–659. https://doi.org/10.1080/00220973.2019.1629577

Şahin, M. D. (2022). Exploratory factor analysis [Açımlayıcı faktör analizi]. In S. Göçer Şahin & M. Buluş (Eds.), *Applied statistics step by step* [*Adım adım uygulamalı istatistik]* (pp. 309–342) Pegem Academy.

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

Temizkan, M., & Sallabaş, M. E. (2015). Comparison of multiple choice tests and open-ended questions in the assessment of reading comprehension skills [Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması]. *Dumlupınar University Journal of Socıal Sciences*, *30*, 207–220.

Terranova, C. (1969). *The effects of negative stems in multiple-choice test items*. Unpublished doctoral dissertation, State University of New York at Buffalo. (30, 2390A).

Vegada, B., Shukla, A., Khilnani, A., Charan, J., & Desai, C. (2016). Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology*, *48*(5), 571–575. https://doi.org/10.4103/0253-7613.190757

Violato, C. (1991). Item difficulty and discrimination as a function of stem completeness. *Psychological Reports*, *69*(3), 739–743. https://doi.org/10.2466/pr0.1991.69.3.739

Violato, C., & Harasym, P. H. (1987). Effects of structural characteristics of stem format of multiple-choice items on item difficulty and discrimination. *Psychological Reports*, *60*(3_part_2), 1259–1262. https://doi.org/10.1177/0033294187060003-251.1

Violato, C., & Marini, A. E. (1989). Effects of stem orientation and completeness of multiple-choice items on item difficulty and discrimination. *Educational and Psychological Measurement*, *49*(1), 287–295. https://doi.org/10.1177/0013164489491032

Yılmaz Koğar, E., & Soysal, S. (2023). Examination of response time effort in TIMSS 2019: Comparison of Singapore and Türkiye. *International Journal of Assessment Tools in Education*, *10*(Special Issue), 174–193. https://doi.org/10.21449/ijate.1343248

Yılmaz, S. (2007). *Misconceptions of second-degree primary school's students about problem solving* (Thesis Number. 200688). [Master Thesis, Eskişehir Osmangazi University], Eskişehir.

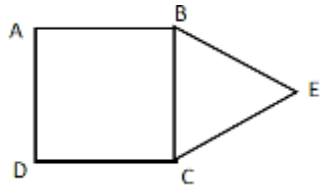
Zhang, F., & Lidbury, B. A. (2013). Evaluating a genetics concepts inventory. In F. Zhang (Eds.), *Sustainable language support practices in science education: Technologies and solutions* (pp. 116–128). Medical Information Scince Reference.

Zulaiha, R., Dian Rahdiani, F., Rahman, A., & Al Anfal, M. F. (2021). Analysis of difficulty level and discriminating power between multiple choices and essay items on math test. *Advances in Social Science, Education and Humanities Research*, *545*, 62–68. https://doi.org/10.2991/assehr.k.210423.065

Appendix

Figure 1

An example of geometry items with and without shapes

The item with shape	The item without shape
 If ABCD is a square and BEC is an equilateral triangle, find the angle of m($\widehat{DCE}$).	ABCD is a square and BEC is an equilateral triangle. If the square ABCD and equilateral triangle BEC have side [BC] in common, find the angle of m($\widehat{DCE}$).
A) 120° B) 130° C) 140° D) 150°	A) 120° B) 130° C) 140° D) 150°

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 209-220

The Efficacy of the IRTree Framework for Detecting Missing Data Mechanisms in Educational Assessments

Yeşim Beril SOĞUKSU*

Abstract

The effectiveness of methods for handling missing data in educational assessments depends on understanding the underlying missing mechanisms. This study investigates the performance of the IRTree framework in detecting missing data mechanisms using a Monte Carlo simulation. Omitted responses were simulated at varying proportions according to three mechanisms: MCAR, MAR, and MNAR, across tests with different lengths and sample sizes. The IRTree was employed to model the omitted responses and detect the mechanisms based on the correlations between the propensity to omit and proficiency. Results indicate that the IRTree accurately identifies all three missing data mechanisms, with no relationship between propensity to omit and proficiency under MCAR, and negative correlations for MAR, reaching up to -0.3, and for MNAR, as high as -0.8. Furthermore, the detection of MAR and MNAR mechanisms became more pronounced with higher proportions of omitted responses, longer tests, and larger sample sizes. IRTree framework not only enables educators and researchers to accurately understand the nature of missing data but also guides them in using appropriate methods for handling it.

Keywords: IRTree, missing data, missing data mechanism, simulation, R language

Introduction

The issue of missing data, which emerges in measurements conducted across various fields such as educational sciences, psychology, healthcare, and social sciences, presents a significant challenge for researchers. Particularly in the fields of education and psychology, where critical decisions about individuals are made, it is essential to understand the nature of missing data and apply appropriate handling methods to ensure that estimates are unbiased and accurate. Missing responses can occur for various reasons across a wide range of testing situations, from classroom to large-scale educational assessments. For instance, in classroom achievement tests, students may omit questions even if they have enough time to answer them due to reasons such as not knowing the answer, difficulty in choosing between options, fatigue, motivation decline or stress. In speed tests, where students must answer as many questions as possible within a given time, it is common for students to leave some items unanswered, particularly towards the end of the test (De Ayala et al., 2001; Graham, 2012; Little & Rubin, 1987). Additionally, in large-scale educational assessments, missing data may result from administering a subset of items to participants to reduce their response burden. This type of planned missing data, known as nonadministered data, typically does not threaten the validity of the assessment (Rose et al., 2015). However, missing responses due to omitted and not-reached items remain one of the most common and problematic issues researchers face during data analysis.

When missing data is not addressed properly, it can lead to biased parameter estimates, Type I and Type II errors, and reduced statistical power. For example, in high-stakes tests where critical decisions are made about students, if ability scores are estimated higher (positive bias) or lower (negative bias) than their actual scores, this bias can result in incorrect decisions about their academic placement or progression. The importance of obtaining unbiased estimates becomes particularly critical in tests where such decisions are based on ability scores. Additionally, errors in hypothesis testing, such as incorrectly

* Dr., Ministry of National Education, Vali Hilmi Tolun Middle School, Kahramanmaraş-Türkiye, e-mail: berilsoguksu@gmail.com, ORCID ID: 0009-0004-0870-4974

To cite this article:
Soğuksu, Y.-B. (2024). The efficacy of the IRTree framework for detecting missing data mechanisms in educational assessments. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 209-220. https://doi.org/10.21031/epod.1514741

Received: 11.07.2024
Accepted: 16.10.2024

rejecting a true null hypothesis (Type I error) or failing to reject the null hypothesis (Type II error), further complicate research findings and diminish the validity of studies (Little et al., 2016; Newman, 2014; Roth, 1994). The importance of unbiased and precise estimates becomes even more evident when considering that countries shape their educational policies based on the results of international assessments like Program for International Student Assessment - PISA (Damiani, 2016; Martens et al., 2016).

The extent to which missing data can affect the validity of measurements depends on their proportions, patterns, and mechanisms (Tabachnick & Fidell, 2007). The proportion of missing responses can significantly impact the generalizability of the study and the statistical inferences drawn from the data. Missing data patterns indicate which responses are observed and which are missing in the dataset but do not provide information about the reasons behind the missing data. Similarly, missing data mechanisms describe the statistical relationships between missing and observed data without explaining how or why the data are missing (Enders, 2010; McKnight et al., 2007).

Rubin (1976) proposed three different mechanisms for missing data: Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In the MCAR mechanism, the presence of missing data is not associated with any variables in the dataset, including the variable itself (Allison, 2002). For instance, in an attitude scale, if students randomly forget to answer some items regardless of their attitudes, this exemplifies the MCAR mechanism. In contrast, the MAR mechanism occurs when the missingness of a variable is related to other observed variables but not to the variable itself (McKnight et al., 2007). For example, in a survey on online learning platforms, students with limited access to technology may have more missing responses. When controlling for other variables, these missing responses are associated with the students' access to technology, not their attitudes towards online learning. The MNAR mechanism occurs when the probability of missing data is directly related to the value of the variable itself (Enders, 2010). An example of this is in competence tests, where lower-performing students tend to leave more questions unanswered, indicating that the missing responses are directly related to their ability levels.

MAR indicates the presence of ignorability, meaning there is no need to model missing data for accurate parameter estimation because the missingness is irrelevant to the parameters being estimated. Many traditional and modern missing methods assume that the missing data are ignorable. When this assumption is not met, it results in biased parameter estimations. Conversely, MNAR data are not ignorable and must be modeled accordingly. Even many modern methods commonly used in the literature can produce biased parameter estimates under the MNAR mechanism (Cheema, 2014; Enders, 2010; Graham, 2012; Rose et al., 2015).

In this context, detecting the missing data mechanisms in a dataset is crucial for researchers to determine the most appropriate methods for handling the missing data. Missing data mechanisms guide the selection of methods that will yield the best performance (Peugh & Enders, 2004). Pigott (2010) emphasized that different methods for handling missing data come with varying assumptions, and if these assumptions are not met, the results can be biased and misleading. For instance, traditional methods often used by researchers assume that missing data meet the MCAR assumption. If this assumption is violated, parameter estimates may be biased, and Type I and Type II errors may occur. In contrast, modern and more robust approaches such as Maximum Likelihood (ML) and Multiple Imputation (MI) operate under the assumption that missing data are ignorable, thereby allowing for unbiased parameter estimates (Graham, 2012; Little et al., 2016; Peugh & Enders, 2004). Therefore, it is essential for researchers to identify the missing data mechanisms in their datasets to apply the most suitable handling techniques.

Several methods can be used to determine whether the missing data mechanism is MCAR. For example, in a scenario where students are surveyed about their math attitudes at the beginning of the semester and their final math exam scores are collected at the end, if students who did not report their attitudes are expected to be no different, on average, from those who did, a t-test can be performed. By comparing the average final math scores between students with missing and complete survey responses, researchers can test for differences. If the missing data mechanism is MCAR, the average math scores should be the same within the sampling error. However, using this method for multiple variables with missing data

can increase the risk of Type I error. Little's MCAR test (Little, 1988) is more effective in such cases, as it helps avoid Type I errors (Enders, 2010).

Little's MCAR test compares the observed variable means for each missing data pattern with the expected population means and calculates a total weighted squared deviation. If the dataset meets the MCAR assumption, any subsample with a given missing data pattern should produce the same means for each variable as those calculated for the entire dataset using a robust parameter estimation method. When there are many patterns of missing data and each pattern tends to produce different means for each variable, the data are unlikely to be MCAR. In this case, the data deviate from a completely random process, and the chi-square test will be significant (McKnight et al., 2007).

Little (1988) contends that MCAR is the only missing data mechanism that can be empirically tested. Similarly, Enders (2010) argues that while methods exist to detect MCAR, it is not feasible to reliably distinguish between MAR and MNAR. Consequently, if the dataset does not meet the MCAR assumption, it is assumed to be either MAR or MNAR. This uncertainty poses a problem because using methods suitable for ignorable missing data on nonignorable missing data can lead to biased parameter estimates. Allison (2002) emphasized that if the MNAR mechanism is incorrectly assumed to be MCAR or MAR, the missing data process will not be modeled accurately, resulting in inaccurate parameter estimates. Similarly, if the MAR mechanism is mistakenly assumed to be MCAR, the estimated parameters will not be generalizable to the population. Huismann (2000) stated that the success of methods used to address missing data depends on accurately identifying the mechanism causing the missing data. In this context, the Item Response Tree (IRTree) framework is recommended as a model-based approach for dealing with missing data, as it provides insights into the missing data mechanisms present in the dataset (De Boeck & Partchev, 2012; Debeer et al., 2017; Jeon & De Boeck, 2016).

IRTree and missing data

The Item Response Tree (IRTree) framework, which integrates item response theory (IRT) and cognitive psychology theories, has garnered significant attention from researchers in recent years (Böckenholt, 2012; Jin et al., 2022). This framework explains response processes through tree-based model structures. There are various studies on the IRTree framework that focus on response styles (Alagöz & Meiser, 2023; Alarcon et al., 2023; Böckenholt, 2017; Dibek, 2019; Plieninger, 2021; Quirk & Kern, 2023; Spratto et al., 2021). Additionally, some studies use the IRTree framework to model answer change behavior (Jeon et al., 2017) and address missing data (Debeer et al., 2017; Huang, 2020; Jeon & De Boeck, 2016). By formulating response probabilities based on tree-based structures, the IRTree method provides a comprehensive framework for understanding response styles while also serving as a robust tool for handling missing data, thereby enhancing the accuracy and validity of parameter estimates in educational and psychological assessments. The ability of the IRTree framework to model non-ignorable missing data, which distinguishes it from both traditional and modern missing data methods, is one of its notable strengths. Unlike commonly used approaches such as Listwise Deletion, which result in a reduction of the dataset, IRTree preserves the integrity of the data. Furthermore, its applicability to both dichotomous and polytomous data, as well as its capability to provide insights into the underlying mechanisms of missing data, positions IRTree as a robust tool for handling missing data in research (Debeer et al., 2017).

The probability of selecting a particular response category in IRTree models depends on the path an individual takes through the tree model. The tree model shown in Figure 1, adapted from Debeer et al. (2017), can be used to model omitted items in a dichotomously scored test. The branching points of this tree model are called ‘nodes,’ and each node represents a different feature based on the underlying assumption of the model.

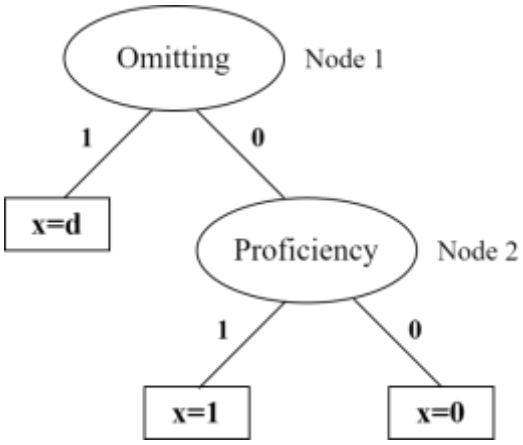


Figure 1. IRTree model for omitted items in dichotomously scored tests

In the tree model shown in Figure 1, when a test taker encounters an item, they first choose one of the behaviors, answer or omit, and then attempts to answer the item correctly. This behavioral process for the test taker is hypothetical. Different tree models can be constructed based on different assumptions. However, it is important that the tree model to be created can logically explain the test-taker's behavioral processes and is appropriate for the test situations. In Figure 1, the tree model has 3 different response categories. These are: Omitting (x=d), correct answer (x=1), incorrect answer (x=0). In addition, there are 2 nodes in this tree model: omitting process and proficiency. The first node, the omitting process, is connected to the proficiency node with two possible response outcomes (two branches from the node), and these nodes can be modeled with different IRT models. Since Birnbaum's (1968) Two Parameter Logistic IRT (2PL) model is used in this study, latent traits for the nodes are $\theta_j^{(1)}$and $\theta_j^{(2)}$ for the person j, slope parameters for item i (i= 1..... k) are $\alpha_i^{(1)}$and $\alpha_i^{(2)}$, and item difficulty parameters for item i are $\beta_i^{(1)}$ and $\beta_i^{(2)}$. This multidimensional model generates the probability of the observed response categories (X_{ij}=0, 1 or d) based on the combination of the probabilities of the sub-processes (Debeer et al., 2017; Jeon & De Boeck, 2016; Park & Wu, 2019).

Since the tree model 1 is a binary model, the right and left branches are coded as 0 and 1, and the observed response categories can be mapped with this coding for the branches. Table 1 shows the mapping matrix for the tree model in Figure 1. If the test taker omits the item in the first stage, the response output for node 1 is 1 and the response output for node 2 is NA; if the test taker answers the item correctly, the response output for node 1 is 0 and is 1 for node 2; if the test taker answers the item incorrectly, the response output for node 1 is 0 and is 0 for node 2. In this way, the data matrix consisting of the test takers' item responses is transformed into a large data matrix containing the responses for nodes 1 and 2. In formulating the probabilities of the response outputs, the response output for each node is modeled with IRT models and the product of these probabilities is taken.

Table 1.
The mapping matrix for omitted items

Original Responses	Node 1	Node 2
Omitted (x=d)	1	NA
Correct (x=1)	0	1
Incorrect (x=0)	0	0

When modeling omitted items in a test, different tree models can be constructed based on different assumptions. At this point, Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to determine which of the different tree models with the same response categories best explains the data structure, thus determining which tree model better fits the data. However, when the tree models have different response categories, it is difficult to compare them directly because the two tree models model different situations. Likelihood ratio tests (LRT) can be used to compare nested tree models (De Boeck & Partchev, 2012; Jeon & De Boeck, 2016).

As a result, after the nodes of tree model in Figure 1 are modeled with the IRT model, the propensity to omit is treated as a latent variable that is different from the measured latent variable of test takers, and the relationship between these two latent variables can be determined (Glas & Pimentel, 2008; Holman & Glas, 2005). At this point, Debeer et al. (2017) state that based on this relationship, it is possible to detect the existing missing data mechanism in a dataset. For example, before choosing the missing data method to be used in a test, determining the extent to which the propensity to omit is related to proficiency using IRTree can guide researchers in terms of using the appropriate handling method. For example, in a competence test if there is a strong negative relationship between these two latent variables (i.e., students with higher abilities omit less items), using traditional missing data methods to deal with missing data will lead to biased results. In this case, the researcher must choose methods that can be used to deal with nonignorable missing data.

When identifying the missing data mechanism in the dataset using the IRTree framework, the correlation between the latent variables for missing propensity and proficiency is examined. While the absence or low correlation between these two latent variables indicates the assumption of MCAR or MAR, a high relationship indicates the presence of nonignorable missing data. In the literature, there are various simulation studies in which different missing data mechanisms are created by manipulating the relationship between these two latent variables. For example, Debeer et al. (2017) generated missing data for MAR based on the absence of a relationship between these two latent variables, and for MNAR based on the relationship being -0.5. Similarly, Huang (2020) generated missing data for MNAR by determining the correlation between these latent variables as -0.5. Holman and Glas (2005) stated that as the correlation between two latent variables increases, the assumption of ignorability weakens. In their study, they showed that when the correlations exceeds 0.4, the nonignorability becomes evident. Köhler et al. (2017) varied the correlation between latent variables as 0.0, 0.2, 0.4, and 0.6, and treated the correlation of 0.0 as MCAR. Glas and Pimentel (2008) varied the correlations between latent variables as 0.0, 0.2, 0.4, 0.6 and 0.8, and stated that the violation of ignorability increases as the correlations move away from 0.0. Glas et al. (2015) interpreted a correlation of 0.0 between latent variables as ignorability, 0.4 as a slight violation of ignorability, and 0.8 as a serious violation of ignorability.

Consequently, the studies in the literature are simulative, and missing data for the MCAR, MAR, and MNAR mechanisms were generated by manipulating the correlations between the missing propensity and proficiency. At this point, a gap exists in the literature, as no study has employed the IRTree framework to identify missing data mechanisms under varying test conditions. Therefore, this study aims to demonstrate the efficacy of IRTree in detecting missing data mechanisms in the presence of omitted items in dichotomously scored tests by examining the correlations between the propensity to omit and proficiency. Specifically, the study explores how the IRTree framework can distinguish between MCAR, MAR, and MNAR mechanisms under varying testing conditions, such as different test lengths, sample sizes, and proportions of omitted responses. After modeling the missing datasets with IRTree, the following research questions were addressed:

- What are the mean correlations between the propensity to omit and proficiency for the missing datasets under MCAR, MAR and MNAR mechanisms?

- Do these mean correlations accurately reflect the underlying missing data mechanisms?

To provide an overview of this paper's structure, the subsequent sections are organized as follows: The Methods section outlines the methodology, including the Monte Carlo simulation used for data generation and the modeling of missing data with IRTree. The Results section provides the results of

the analyses, focusing on how well the IRTree framework was able to detect the missing data mechanisms under different test conditions. Finally, the Discussion section discusses the implications of these findings in the context of existing research and highlights the contributions of this study.

Methods

Data Generation

The study was conducted as a Monte Carlo simulation, using the mirt package (Chalmers, 2012) in R to generate the datasets. Data generation was performed according to Birnbaum's (1968) Two-Parameter Logistic (2PL) model. The formula for the 2PL model is as follows:

$$P(\theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (1)$$

In the Formula 1, θ_j refers to the latent trait of the person j, a_i refers to the item discrimination parameter, b_i refers to the item difficulty parameter for item i. While selecting the distributions of item and ability parameters, the previous studies were considered (Baker, 2001; DeMars, 2010; Feinberg & Rubright, 2016; Hambleton et al., 1991; Harwell et al., 1996). In this context, the discrimination parameters follow a uniform distribution a ~ U(0, 2). Item difficulty parameters and ability parameters were drawn from the normal distribution θ ~ N(0, 1). To avoid convergence issues and ensure accurate parameter estimation, test lengths were set at 20 and 40 items (Alarcon et al., 2023; DeMars, 2010). As in previous studies (Alarcon et al., 2023; Leventhal, 2019), a sample size of 500 was included in this study to assess the robustness of the IRTree framework with small sample sizes, which is commonly encountered in educational assessments. In addition, sample sizes of 1000 and 2000 were chosen based on previous studies on modeling missing data with IRTree (Debeer et al., 2017; Huang, 2020). Since it is necessary to have at least 5% omitted responses to effectively model the omitting process with IRTree (Debeer et al., 2017), the lowest missing data proportion in the study was set at 5%. Proportions of 10%, 30%, and 50% were also chosen to create test scenarios with varying levels of missing data and assess the robustness of the IRTree. The number of replications was set at 100. The study was conducted using the R programming language, version 4.3.2.

Types of missing data

After data generation, missing data were created for three different missing data mechanisms. For MCAR, %5, 10%, 30%, and 50% of the responses were randomly deleted from the datasets. For MAR, following the approach of Collins, Schafer, and Kam (2001), a covariate variable was generated that correlated with the total test scores, with the correlation set at 0.3. The covariate variable was divided into three group based on its quartiles (.00 - .33, .33 - .66, .66 - 1.00), and different probabilities of missing data were assigned to each group. At this point, test takers in the lower quartile group have a higher probability of missing data, whereas test takers in the upper quartile group have a lower probability of missing data. For example, in the 10% missing data scenario, 15% of observations were randomly deleted from group 1, 10% from group 2, and 5% from group 3, ensuring that the mean missing data proportion was 10%. For MNAR, as used in previous studies (Rose et al., 2010; Sulis & Porcu, 2017), the ability scores θ were considered. The ability scores were estimated and each sample sizes are divided into three groups based on the ability scores' quartiles (.00 - .33, .33 - .66, .66 - 1.00). Missing data were generated such that test takers with higher ability levels had a lower probability of missing data, while those with lower ability levels had a higher probability. For instance, in the 10% missing data scenario, 15% of observations were randomly deleted from the group with the lowest ability level, 10% from the middle group, and 5% from the highest ability group, ensuring a mean data proportion of 10%.

Analysis

Missing data sets were modeled using the tree model for omitted items in Figure 1 proposed by Debeer et al. (2017). This model includes two nodes: propensity to omit and proficiency. The original responses in the datasets were converted into mapping matrices based on this tree model using the dendrify2 function from the flirt package (Jeon, Rijmen, & Rabe-Hesketh, 2014). Subsequently, each node in the tree model was modeled with the 2PL model using the mirt function (Chalmers, 2012) in the wide data matrix format. Latent variables for the nodes were estimated using the Expected a Posteriori (EAP) method. The EAP method, proposed by Bock and Aitkin (1981), calculates the mean of the ability parameter distribution given the observed response pattern. Unlike Maximum Likelihood (ML) estimations, EAP estimations can be computed even when a test taker answers all items correctly or incorrectly (Bock & Mislevy, 1982; De Ayala et al., 2001). Since the missing data mechanisms were determined based on the correlation between the propensity to omit and proficiency, the mean Pearson correlation coefficient was calculated for all conditions in the study. If the mean correlation between these two latent variables is greater than 0.4, it is considered a violation of ignorability as in the literature (Debeer et al., 2017; Glas et al., 2015; Holman & Glas, 2005; Huang, 2020).

Results

In this study, the effectiveness of IRTree in detecting missing data mechanisms was tested under different simulation conditions. The correlations between the propensity to omit and proficiency were calculated, and the mean of these correlations was then computed. Figure 2 illustrates the mean correlations for MCAR mechanism across the various conditions.

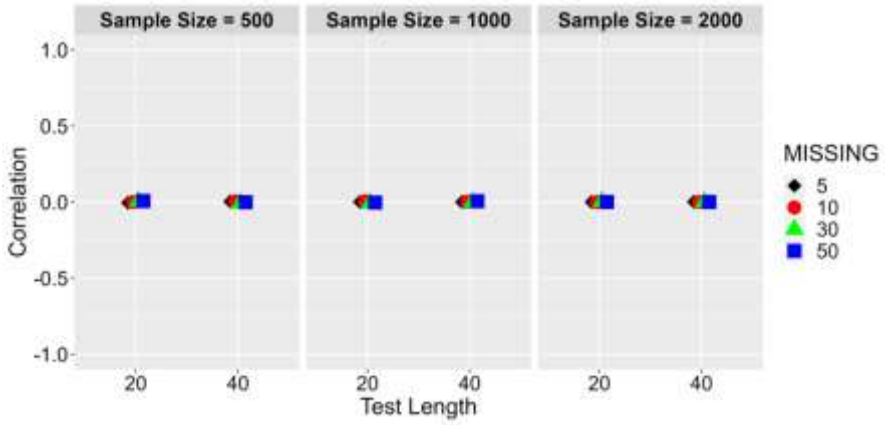


Figure 2. Mean correlations between propensity to omit and proficiency for MCAR

As shown in Figure 2, for MCAR, the mean correlations between the propensity to omit and proficiency are 0.0 for each condition. Mean correlations remained consistent across varying test lengths, sample sizes, and missing proportions. This indicates that under all conditions, there is no relationship between test takers' proficiency and propensity to omit, confirming that the missing data do not contain information about the measured latent trait and that the missingness is unrelated to both observed and unobserved variables.

The missing data for MCAR were generated by randomly deleting specified proportions of responses from the datasets. As a result, no relationship was expected between test takers' proficiency and their propensity to omit. Based on the results obtained through the IRTree framework, it can be concluded that the MCAR mechanism was correctly detected. Figure 3 illustrates the mean correlations across the conditions for MAR.

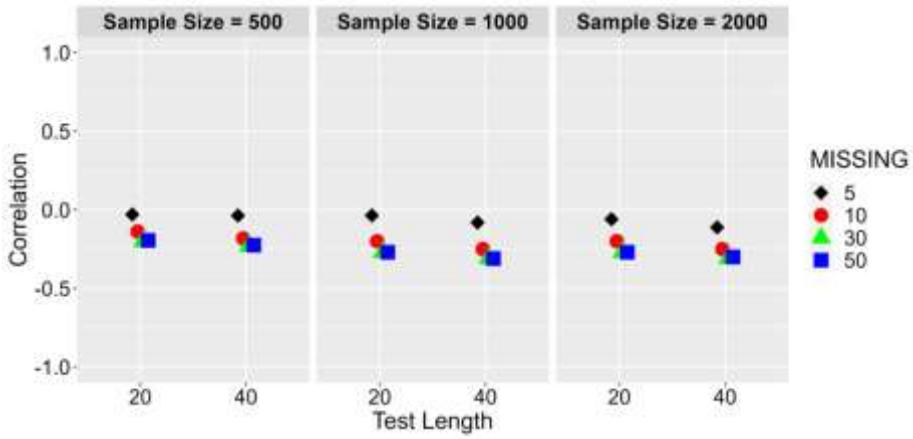


Figure 3. Mean correlations between propensity to omit and proficiency for MAR

In Figure 3, the mean correlations between the propensity to omit and proficiency are negative and up to -0.3. The negative mean correlations suggest that as test takers' proficiency increases, their propensity to omit decreases. However, the weakness of the relationship indicates that there is no violation of ignorability and that the missingness is not strongly related to proficiency.

Moreover, it was observed that an increase in the missing data proportion strengthens the relationship between the propensity to omit and proficiency. At this point, with a 5% missing data proportion, the missing data mechanism appears to align with MCAR, particularly in cases with smaller sample sizes and shorter tests. Therefore, for the relationship between propensity to omit and proficiency to adequately reflect the MAR assumption, and for the IRTree to effectively capture this relationship, the dataset should ideally contain at least 10% omitted responses. The increase in test length led to a slight rise in mean correlations, while increasing the sample size from 500 to 1000 resulted in higher mean correlations. However, increasing the sample size from 1000 to 2000 did not have an impact on the correlations, particularly when the missing data proportion exceeded 5%.

To generate missing data for MAR, a covariate variable with a low correlation to the total scores was created, and missing data were generated in increasing proportions based on the quartiles of this covariate variable. In this scenario, the probability of missing data does not depend on the test takers' proficiency, but rather on another (covariate) variable in the data set. Consequently, in the IRTree modeling, the low mean correlations between propensity to omit and proficiency successfully reflect this situation. Figure 4 illustrates the mean correlations across the conditions for MNAR.

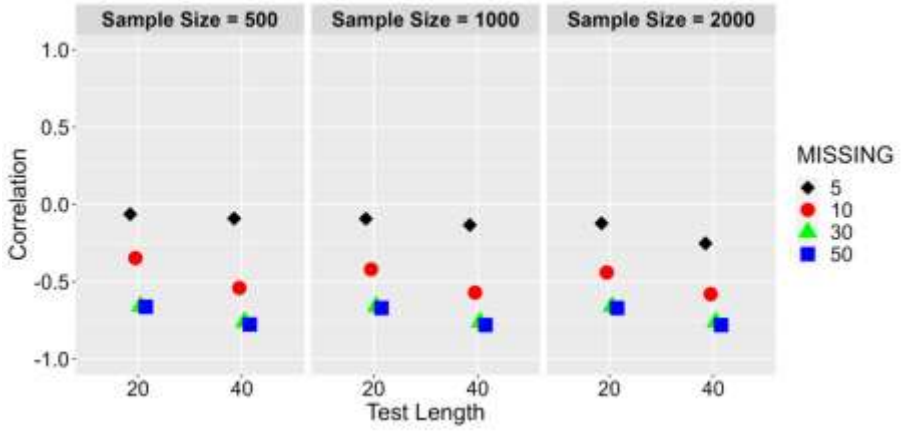


Figure 4. Mean correlations between propensity to omit and proficiency for MNAR

As shown in Figure 4, the mean correlations are negative and up to -0.8. However, when the missing data proportion is 5%, the relationship between propensity to omit and proficiency approaches 0.0, indicating either an MCAR or MAR mechanism. Nevertheless, as the missing data proportion increases, the mean correlations rise from moderate to high, indicating that the propensity to omit decreases as test takers' proficiency increases. This is consistent with the expectation that test takers with higher ability levels would omit fewer items, reflecting the MNAR data generation process. Therefore, the missing data contain information about proficiency and cannot be ignored.

The increase in the missing data proportion leads to a negative increase in the mean correlations, strengthening the MNAR mechanism. At this point, for the MNAR mechanism to be effectively detected using the IRTree, it is recommended that the dataset contain at least 10% omitted responses. Additionally, the mean correlations increased negatively with the increase in test length. While the increase in sample size resulted in higher mean correlations at 5% and 10% missing data proportions, changes in sample size did not cause a noticeable difference in mean correlations at 30% and 50% missing data proportions. At this point, after reaching a 30% missing data proportion, further increases in sample size did not affect the mean correlations.

While generating missing data for MNAR, the probability of missing data was associated with the ability of the test takers. High-ability test takers were less likely to omit items, whereas low-ability test takers were more likely to do so. Consequently, the missing data contained information about the latent trait being measured. In the IRTree modeling, the negative, moderate to high correlations between missing propensity and proficiency, especially in cases with 10% or more omitted responses, successfully reflect this situation.

Discussion

To ensure accurate assessments and evaluations in educational settings, it is important to effectively handle missing data. Understanding the missing data mechanism present in the dataset is crucial to using appropriate handling methods that avoid bias and error in parameter estimations. The literature shows that various methods can identify MCAR, but distinguishing between MAR and MNAR is challenging (Enders, 2010; McKnight et al., 2007). Consequently, researchers often cannot determine whether the missing data meet the MAR or MNAR assumption, potentially leading to biased parameter estimates. This study aimed to evaluate the efficacy of the IRTree framework in detecting missing data mechanisms under different test conditions. Specifically, it focused on items that respondents omitted for various reasons in dichotomously scored tests. Using a Monte Carlo simulation, datasets with missing data under MCAR, MAR, and MNAR were modeled with the IRTree model for omitted items. The mean Pearson correlation coefficients between the propensity to omit and proficiency were examined to detect the missing data mechanisms.

The IRTree modeling revealed no correlation between propensity to omit and proficiency under the MCAR mechanism, with mean correlations of 0.0 as observed in the literature (Glas & Pimentel, 2008; Köhler et al., 2017). For MAR, the mean correlations were negative, reaching up to -0.3, indicating that highly proficient test takers were less likely to omit items, though this relationship is weak. According to Holman and Glas (2005), nonignorability becomes evident when the correlation exceeds 0.4. Since the observed correlations in MAR remain below this threshold, the data can be considered ignorable. For MNAR, mean correlations were as high as -0.8, especially at the highest missing data proportion, indicating that the missing data process contains information about the proficiency. As test takers' proficiency increases, their propensity to omit decreases. This result is consistent with the finding that as the correlation between propensity to omit and proficiency increases, the level of nonignorability also rises (Glas & Pimentel, 2008; Pohl et al., 2014).

Varying missing data proportions, test lengths, and sample sizes did not affect the relationship between propensity to omit and proficiency for the MCAR mechanism, and the IRTree was able to detect MCAR under all conditions. For MAR and MNAR, however, the relationship between propensity to omit and proficiency increased with the rising proportion of missing data, with a more pronounced impact in

MNAR. Additionally, for the relationship between propensity to omit and proficiency to reflect MAR and MNAR, the dataset should ideally contain at least 10% omitted responses. Increasing the number of items raised the mean correlations in both MAR and MNAR mechanisms, with a more pronounced impact in MNAR. For MAR and MNAR, increasing the sample size from 500 to 1000 led to an increase in the mean correlations with the effect being more pronounced for MNAR. While further increasing the sample size from 1000 to 2000 did not result in a substantial change. Consequently, increases in the proportion of omitted responses, test length and sample size strengthened the relationship between the propensity to omit and proficiency, especially making the MNAR mechanism more pronounced and easier to detect.

Overall, the study demonstrated that the IRTree framework could accurately detect missing data mechanisms across various scenarios with different sample sizes, test lengths, and missing proportions in dichotomously scored tests. This finding supports the utility of IRTree in identifying missing data mechanisms, as suggested by previous studies (Debeer et al., 2017; Jeon & De Boeck, 2016). Through IRTree analyses, it is possible to determine whether the measured latent traits are related to the process that leads to missing data, enabling more informed and rational decisions on how to handle omitted items in real-world educational assessments. For example, if omitting behavior has a high correlation with the measured latent trait, in such a case, the Selection Model, Pattern Mixture Model or IRTree can be used to deal with nonignorable missing data will need to be employed (Debeer et al., 2017; Enders, 2010; Holman and Glas, 2005). This will increase the accuracy of the results by preventing biased parameter estimates. Accurate and precise estimates will enhance the validity and reliability of assessments, preventing incorrect decisions about test takers. Especially in tests that measure students' achievements or attitudes, using IRTree to determine whether omission behavior is related to the measured trait will guide educators in their decision-making processes. Additionally, in large-scale educational assessments such as PISA, TIMMS, or PIRLS, which guide countries' educational policies, IRTree results can provide a clearer understanding of the nature of missing data.

This study focused on omitted responses in dichotomously scored tests. Future research can extend this work by investigating the effectiveness of the IRTree framework in detecting missing data mechanisms in polytomously scored tests. Additionally, future studies should test the efficacy of IRTree under different simulation scenarios. Different IRTtree models can be developed by considering various processes leading to missing data. While this study used the 2PL model for the nodes of the IRTree model, future research can explore the use of different IRT models.

Declarations

Conflict of Interest: The author reports there are no competing interests to declare.

Ethical Approval: I declare that all ethical guidelines for the author have been followed. Ethical approval is not required as data has been simulated in this study.

Funding: The author received no financial support for the research, authorship, and/or publication of this article.

References

Alagöz, Ö. E. C., & Meiser, T. (2023). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement, 84*(5), 957-993. https://doi.org/10.1177/00131644231206765

Alarcon, G. M., Lee, M. A., & Johnson, D. (2023). A Monte Carlo study of IRTree models' ability to recover item parameters. *Frontiers In Psychology, 14*, 1003756. https://doi.org/10.3389/fpsyg.2023.1003756

Allison, P. D. (2002). *Missing data.* Sage Publications.

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

Bock, R. D., & Aitkin M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444. https://doi.org/10.1177/014662168200600405

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-460). MA: Addison-Wesley.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665-678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological methods, 22*(1), 69–83. https://doi.org/10.1037/met0000106

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research, 84*(4), 487-508. https://doi.org/10.3102/0034654314532697

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-51.

Damiani, V. (2016). Large-scale assessments and educational policies in Italy. *Research Papers in Education, 31*(5), 529–541.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*(3), 213–234. https://doi.org/10.1111/j.1745-3984.2001.tb01124.x

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1-28. https://doi.org/10.18637/jss.v048.c01

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement, 54*(3), 333-363. https://doi.org/10.1111/jedm.12147

DeMars, C. (2010). *Item response theory: Understanding statistics measurement.* Oxford University Press.

Dibek, M. I. (2019). Examination of the extreme response style of students using IRTree: The case of TIMMS 2015. *International Journal of Assessment Tools in Education, 6*, 300-313. https://doi.org/10.21449/ijate.534118

Enders, C. K. (2010). *Applied missing data analysis.* The Guilford Press.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. https://doi.org/10.1111/emip.12111

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 48*(6), 907-922. https://doi.org/10.1177/0013164408315262

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous models with covariates. *Psychological Test and Assessment Modeling, 57*(4), 523-541.

Graham, J. W. (2012). *Missing data analysis and design.* Springer.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. California: Sage Publications.

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Pyschology, 58*, 1-17. https://doi.org/10.1111/j.2044-8317.2005.tb00312.x

Huang, H. Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement, 80*(6), 1168-1195. https://doi.org/10.1177/0013164420914711

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity, 34*, 331–351. https://doi.org/10.1023/A:1004782230065

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*, 1070-1085. https://doi.org/10.3758/s13428-015-0631-y

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics, 42*(4), 467-490. https://doi.org/10.3102/1076998616688015

Jeon, M., Rijmen, F. & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. Applied Psychological Measurement, 38, 404-405. https://doi.org/10.1177/0146621614524982

Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2022). A new multiprocess IRT model with ideal points for likert-type items. *Journal of Educational and Behavioral Statistics, 47*(3), 297-321. https://doi.org/10.3102/10769986211057160

Köhler, C., Pohl, S., & Carstensen, C. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement, 54*, 397-419. https://doi.org/10.1111/jedm.12154

Leventhal, B. C. (2019). Extreme response style: A simulation study comparison of three multidimensional item response models. *Applied Psychological Measurement, 43*(4), 322–335. https://doi.org/10.1177/0146621618789392

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* John Wiley & Sons.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202.

Little, T. D., Lang, K. M., Wu, W., & Rhemtulla, M. (2016). Developmental psychopathology. In D. Cicchetti (Ed.), *Missing Data* (pp. 760-797). John Wiley & Sons.

Martens, K., Niemann, D., & Teltemann, J. (2016). Effects of international assessments in education – a multidisciplinary review. *European Educational Research Journal, 15*(5), 516-522. https://doi.org/10.1177/1474904116668886

McKnight, P. E., McKnight, K. M., Sidani, S. & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* Guilford Press.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational research methods, 17*(4), 372-411. https://doi.org/10.1177/1094428114548590

Park, M., & Wu, A. D. (2019). Item response tree models to investigate acquiescence and extreme response styles in Likert-type rating scales. Educational and Psychological Measurement, 79(5), 911–930. https://doi.org/10.1177/0013164419829855

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525-556. https://doi.org/10.3102/00346543074004525

Pigott, T. D. (2010). A review of methods for missing data. *Educational Research and Evaluation: An International Journal on Theory and Practice, 7*(4), 353-383. https://doi.org/10.1076/edre.7.4.353.8937

Plieninger, H. (2021). Developing and applying Ir-Tree models: Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods, 24*(3), 654-670. https://doi.org/10.1177/1094428120911096

Pohl, S., Gräfe, L. & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. Educational and Psychological Measurement, 74(3), 423–452. https://doi.org/10.1177/0013164413504926

Quirk, V. L., & Kern, J. L. (2023). Using IRTree models to promote selection validity in the presence of extreme response styles. *Journal of Intelligence, 11*(11), 216. https://doi.org/10.3390/jintelligence11110216

Rose, N., von Davier, M., & Nagengast, B. (2015). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82*, 795-819. https://doi.org/10.1007/s11336-016-9544-7

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (ETS Research Report No. RR-10-11). Educational Testing Service.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Spratto, E. M., Leventhal, B. C., & Bandalos, D. L. (2021). Seeing the forest and the trees: Comparison of two IRTree models to investigate the impact of full versus endpoint-only response option labeling. *Educational and Psychological Measurement, 81*(1), 39-60. https://doi.org/10.1177/0013164420918655

Sulis, I., & Porcu, M. (2017). Handling missing data in item response theory. Assessing the accuracy of a multiple imputation procedure based on latent class analysis. *Journal of Classification, 34*, 327–359. https://doi.org/10.1007/s00357-017-9220-3

Tabachnick, B. G., & Fidell L. S. (2007). *Using multivariate statistics.* Allyn and Bacon.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 221-246

Meta-Analytic Reliability Generalization Study of Perceived Stress Scale in Türkiye Sample

Ömer DOĞAN* Selahattin GELBAL**

Abstract

The aim of the study is to examine the meta-analytic reliability generalization of the 14, 10, and 8-item forms of the Perceived Stress Scale, which was developed by Cohen, Kamarck and Mermelstein in 1983 and translated into Turkish by different researchers (Erci, 2006; Yerlikaya & İnanç, 2007; Eskin et al. 2013, etc.) between 2006 and 2013, for the theses produced in Türkiye. For this purpose, how different moderator variables affect the reliability coefficients and publication bias were also examined. A total of 81 Cronbach Alpha coefficients from 78 studies, selected in accordance with the established criteria, were included in the meta-analysis. Reliability generalization was applied in the data analysis. The analyses were conducted using the random effects model with coefficient values converted through the Bonett method. In the study, the effect size value was found to be .82 (95% CI: .80, .83), and it was concluded that the sample type (α= .81) and the study area (α= .81) moderator variables had a statistically significant effect on the reliability estimation in terms of heterogeneity of effect sizes. This revealed that these two moderator variables affected the estimations of the reliability coefficients. In addition, it has been observed that other moderator variables such as age, gender, and the number of items in the scale are not sources of heterogeneity and have no effect on reliability estimation. From this, it was deduced that the scale works well enough to generalize to different contexts with different item numbers. Finally, according to the results of the analysis for the determination of publication bias, it was seen that there was no publication bias in the study.

Keywords: reliability generalization, meta-analysis, perceived stress scale, Cronbach alpha coefficient, Bonett method.

Introduction

One of the important reasons for the rapid progress of science is that it is cumulative. Accumulation of science is valuable with the accuracy of the inferences that are considered scientific knowledge. One of the conditions required for knowledge to be scientific is to obtain that knowledge by measuring it with the help of tools whose standards have been determined. The usefulness of the measurements depends on the robustness of the qualifications of the measurement tool. The most important of these qualities are validity and reliability. Validity evidence and reliability coefficients of the scores obtained from the measurement tools should be examined. One of Cronbach's (1947) definitions for reliability among these concepts is the degree to which the test score reflects the individual's current state in the general and group factors defined by the test. Anastasi (1976) defines reliability as the consistency of scores obtained when re-examined by the same people in different situations with the same test or with different sets of equivalent items or under other varying conditions. In the Standards (American Educational Research Association (AERA) et al., 2014), reliability is defined as the correlation between the scores on two equivalent forms of the test and is used to refer to the consistency of scores across replications of a test procedure.

As stated in Classical Test Theory, reliability is not an inherent property of the test but a value derived from a specific implementation. As with validity, test score reliability should be considered in relation to specific test purposes and contexts. Defining, measuring and reporting reliability should begin with considerations of the intended uses and interpretations of the test. The reliability of test scores varies

* Dr. Ministry of National Education, Usak- Türkiye, e-mail: 64omerdogan64@gmail.com, ORCID ID: 0000-0001-5169-520X
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, e-mail: gelbal@hacettepe.edu.tr, ORCID ID: 0000-0001-5181-7262

To cite this article:
Doğan, Ö. & Gelbal, S. (2024). Meta-analytic reliability generalization study of perceived stress scale in Türkiye sample. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 221-246. https://doi.org/10.21031/epod.1536530

Received: 20.08.2024
Accepted: 17.10.2024

according to the sample structure, diversity and context of implementation (Crocker & Algina, 1986; Haertel, 2006; Streiner et al., 2015). Therefore, many scientific institutions, journals, and researchers have recommended reporting the reliability estimates with the data obtained from the studies and avoiding the misuse of reliability coefficients from previous test applications (Vacha-Haase et al., 2000; Wilkinson, L., & the APA Task Force on Statistical Inference, 1999). Cronbach's Alpha is the most commonly used coefficient in the literature to analyze and interpret internal consistency reliability (Özdemir et al., 2020; Urbina, 2004). Nevertheless, as in all coefficients, Cronbach's Alpha coefficient varies from study to study even if the same scale is used since it is a sample-dependent coefficient. For example, while the Cronbach Alpha value was .60 in one study (Güler, M.Ş., 2019) using the Perceived Stress Scale (PSS), it was found to be .97 in another study (Güler, B., 2019).

Reliability analyses depend on the areas of variability allowed in the test procedure (e.g., tasks, contexts, raters) and the proposed interpretation of test scores (AERA et al., 2014). In contrast to common misconceptions, reliability is a dynamic characteristic of test scores rather than a fixed value for measurement results and can vary according to the properties of the data (Thompson & Vacha-Haase, 2000 as cited in Eser & Doğan, 2023). Therefore, reliability should be calculated again after any measurement and reported in each study. Reliability coefficients may also differ depending on the variety of sample characteristics, and for this reason, sample size, scale implementation conditions, implementation time etc. differences require the generalization of the reliability of the studies conducted. Reliability generalization (RG) was initially developed by Vacha-Haase (1998). RG analyzes the sources and amount of variability of reliability coefficients in different studies (Vacha-Haase, 1998). RG studies are subsequent analyses aimed at establishing of an average estimate of the effect sizes observed across studies, and specifically to search for evidence of reliability. (Borenstein, 2009; Hunter & Schmidt, 2004; Vacha-Hasse, 1998 as cited in Allan, 2021).

RG studies are relatively new, but they provide beneficial information beyond the simple definition of score reliability for an instrument. A properly conducted RG study can provide more accurate estimates of score reliability by synthesizing reliability estimates from a data set of estimates from individual samples (Beretvas et al., 2008).There are various sources and studies in the literature related to RG studies. (e.g., Aguayo et al., 2011; Allan, 2021; Alzahrani, 2016; Hongyan et al., 2015; Nicolas et al., 2021). However, only Eser and Aksu (2021), Özdemir et al. (2020), Şen (2021a) and Şeten (2012) studies are available in the Turkish literature. It is seen that the interest in RG study has been increasing in recent years. Unfortunately, however, as many RG studies reveal, there is still a critical shortage of studies reporting reliability estimates for the samples used. This is problematic for several causes. First, underreporting continues to spread the misconception that reliability is a property of the scale, not the scores derived from it. Second, under-reporting limits the data that can be used in an RG study to provide a precise and accurate pooled estimate of reliability (Beretvas et al., 2008). In RG studies, the moderator variables that affect the reliability coefficient are generally handled as sample size, sample type, year of study, participant characteristics, etc. This study investigated the generalizability of reliability coefficients based on sample size, gender, year of study, number of items in the scale, sample type, field of study, thesis type and average age of the participants, and whether the reliability coefficients are affected by these variables. The reasons for selecting these variables are that perceived stress varies according to gender (Graves et al. 2021), age (Osmanovic-Thornström et al., 2015), sample type and the job (Lee et al., 2012) in the context of the field of study. However, the desire to examine the effect of sample size, thesis type and number of items that affect study quality and reliability in the context of the studies conducted required the inclusion of these variables in the study. In addition, since the number of studies using the perceived stress scale was more than 500, a limitation was necessary. For this reason, theses that are believed to have been prepared more carefully and in which the number of samples is generally higher were considered within the scope of the research. Even the number of theses written on this subject is well above the number of studies used in many RG studies.

Within the scope of the study, it was aimed to examine the reliability generalization of the PSS developed by Cohen et al. (1983) in forms with different number of items. The PSS, developed by Cohen et al. (1983), was chosen because it is the most widely used stress scale. In addition, within the scope of scale development, it was found important that three different samples were formed and analyzed, and

the reliability coefficients obtained from these samples were high, with values of .84, .85, and .86. In addition, the fact that there was no significant difference in the comparison of male and female participants in the samples provides the possibility of applying the scale to all individuals. Stress has become one of the most commonly used concepts in daily life in recent years. Stress has become one of the most commonly used concepts in daily life in recent years. It is often used to mean "anxiety," "worry," and "tension," and is defined as an external load or demand on a mental, biological, social , or psychological system (Lazarus, 1993). Following the definition of stress, theoretical frameworks, such as Lazarus' Stress and Related Relationships Theory, have been developed to provide a comprehensive context for how it is measured. Questions such as what stress is, its sources, its physical and psychological symptoms, when it becomes dangerous, and how to manage it have become important. Although people have similar experiences, stress responses and outcomes may differ (Cohen et al., 1997; Lazarus & Launier, 1978). This suggests that people differ in the way they interpret and react to events. Questions regarding what stress is, its sources, physical and psychological symptoms, when it becomes dangerous, and how to manage it have gained importance. Although people have similar experiences, stress reactions and outcomes may vary (Cohen et al., 1997; Lazarus & Launier, 1978); that is, people differ in the way they interpret and react to situations. Stress has a negative impact on people's ordinary actions in their lives and on their quality of life. Lazarus (1990) discussed four controversial issues related to stress and adopted the view that stress is a subjective rather than objective phenomenon, that it can be measured as relatively minor adversity rather than major catastrophes, that the relationship between stress and adaptive outcomes is not due to confounding, and that any measure of stress should assess the content or sources of stress rather than its degree. He also advocated for a greater emphasis on the psychological content of stress scales and more attention to the broader adaptive context of the individual, the systems theory perspective, and the time periods over which stress is measured. Among the types and definitions of stress, perceived stress represents a global and comprehensive construct of stress and is based on the concept that individuals actively interact with their environment, evaluating potentially threatening or challenging events considering available coping resources (Katsarou et al., 2013). Studies have shown that the increase in the perceived stress levels of individuals negatively affects their quality of work and life (Camci, 2021; Havare, 2019; Ataman Temizel & Dağ, 2014 etc.). Cohen (1994) stated that the Perceived Stress Scale (PSS) is one of the most common psychological instruments used to measure stress perception and that the scale is a measure of the extent to which situations in one's life are considered stressful. The scale items are constructed to measure how unpredictable, uncontrollable and overloaded individuals find their lives. The scale also includes a series of direct questions about current stress levels. Within the scope of the research, Cronbach Alpha coefficient was used as the reliability coefficient. There are areas where the Cronbach Alpha coefficient is incomplete (Agbo, 2010), and it is sometimes misused by researchers. Although these are some of the reasons why more than one reliability coefficient is recommended (Agbo, 2010), Cronbach's Alpha is the most widely used reliability measure (Hussey et al., 2023) and used in almost all studies employing this scale. Yerlikaya and İnanç (2007) adapted the 14-item form of the PSS into Turkish. Some items in the scale (4th, 5th, 6th, 7th, 9th, 10th and 13th) were reverse coded and the internal consistency coefficient of the scale was found to be 0.84. In the adaptation made by Eskin et al. (2013) for the 14-item form of the PSS, the internal consistency coefficient was found to be 0.86. The first of the two factors in the scale was named as "Insufficient Self-Efficacy Perception" and the other as "Stress/Discomfort Perception". It was stated that these two factors explained 46.5% of the total variance. The Turkish adaptation study of the 10-item form of PSS was also carried out by different researchers. In the first Turkish adaptation study conducted by Erci (2006), 4 items were scored as positive and 6 items as negative. The Alpha coefficient of the scale was found to be 0.70 and it was stated that the scale explained 58.1% of the total variance. The internal consistency coefficient of the PSS-10 scale, adapted to Turkish by Çelik Örücü and Demir (2009), was found to be 0.84. In Eskin et al. (2013), it was found that the scale consisted of two factors: "insufficient self-efficacy" and "perception of stress or discomfort". The reliability values of the two factors were determined as .69 and .80. The total reliability of the scale was found to be .82. The 8-item form of the PSS was adapted by Bilge et al. (2009) and prepared in a 5-point Likert type (0 never, 4 very often) and three items were scored in reverse (4th, 5th, 6th). In addition, the Cronbach Alpha coefficient was found to be .81. The reliability coefficients of the PSS scale in different cultures are also quite high. For example, in a study

conducted in Mexico, Ramírez and Hernández (2007) found the Cronbach Alpha coefficient to be .83, and in a study conducted in Greece, Andreou et al. (2011) found the Cronbach Alpha coefficient to be .82. The reason for choosing the PSS in this study is that stress, defined as physical strain that can cause physical and mental diseases by reducing the body's resistance (Birol & Akdemir, 2003), is associated with many concepts in various fields. In addition, in recent years, especially during the pandemic period, the number of studies on the concept of stress has increased in parallel with the increase in people's stress levels and types. The Coronavirus Disease-2019 (COVID -19) pandemic has caused significant stress for humanity (Singh et al. 2021). A study by Qui et al. (2020) showed that stress was high among young people during the pandemic. However, another study by Pedrozo-Pupo et al. (2020) showed that stress was found to be similar across all age groups. The young population struggled with stressors such as the absence of face-to-face academic, physical and social activities, disruption of routine and boredom. Adults struggled with specific issues such as financial burdens, social isolation, the unpredictability of life and increased childcare responsibilities due to school closures (Gallagher et al. 2020). Considering that stress has a significant impact on people's life and health, it is extremely important to reliably measure the level of perceived stress. An examination of both Turkish and international literature showed that the PSS was frequently used to measure perceived stress levels. Since the number of studies conducted in this field is very large, the scope of the research is limited to the theses in which PSS is used in the Turkish sample. 31 of the 164 theses, in which PSS was used in the process until the end of the study, belong to the year 2021 only. Considering that the reliability values of the studies using the PSS in the theses examined in Turkish literature range from .60 to .97, it would be useful to investigate these differences and generalize the reliability findings to the Turkish sample.

Given that there are different perceived stress scales (e.g., for prenatal pregnant women (Razurel et al., 2014), for nursing students (Sheu et al., 2002), for children (Snoeren & Hoefnagels, 2014), it is thought that a good understanding of the psychometric properties of the PSS concept by researchers and an RG of the PSS scale would provide an important basis for researchers to consider the PSS in future studies. Therefore, the aim of the study was to conduct an RG meta-analysis to investigate the variability of PSS score reliability across studies. RG studies can help to understand how reliable the scores produced by the scale are across different samples, number of items, and fields of study, etc., rather than whether the measurement is reliable or not. It is thought by Vacha-Haase et al. (2002) that RG studies can contribute to field experts and individuals who will develop scales. As a result of all these, it is essential that this study be included in the Turkish literature and in the fields of education, health and psychology. For that reason, a meta-analytic RG analysis was conducted for the Turkish sample of the PSS and it was examined how the reliability coefficients were generalized according to the moderator variables of number of items, sample type, field of study, sample age, year of publication, type of thesis, gender and sample size, and whether the reliability coefficients were affected by these variables. For this reason, a meta-analytic RG analysis was conducted for the Turkish sample of the PSS, examining how the reliability coefficients generalized according to the moderator variables of the number of items, sample type, field of study, sample age, publication year, thesis type, gender, and sample size, and whether these variables affected the reliability coefficients. Finally, the question of whether the reliability coefficients were affected by these variables was sought to be answered. The results of this study will inform researchers who plan to conduct studies using this scale about the range of reliability estimates that can be expected for the PSS and ensure that sources of variability are considered.

Methods

In this research, the meta-analysis method was used to synthesize the results of these studies by considering and examining them for a specific purpose (Büyüköztürk et al., 2019). This section includes sampling, coding of study variables, and data analysis. The RG study was conducted according to the PRISMA (Liberati et al., 2009) guidelines. Accordingly, the databases of the theses were searched and the studies suitable for the criteria were determined. Then, identical studies were removed and studies were evaluated for inclusion in the meta-analysis.

Data Collection Process

The selection process for the studies included in the research is presented in Figure 1 (Liberati et al., 2009). Figure 1 illustrates that the first step is to search for relevant studies. As mentioned above, since the number of studies conducted using PSS is very high, it was decided to examine only the theses in which PSS was used and the theses were scanned from https://tez.yok.gov.tr/UlusalTezMerkezi/ where the theses are uploaded in Türkiye. The keywords used were "perceived stress" and "stress they perceive". When it was searched with keywords on the specified site, it was seen that there were a total of 164 studies. First of all, double coding was avoided for the theses reached by using both keywords. In addition, for studies containing two or more reliability coefficients, the coefficients were coded independently. Therefore, a total of 157 studies out of 164 studies were coded. These 157 studies were then reviewed according to the inclusion criteria. The criteria were as follows:

I) Cronbach Alpha reliability coefficients should be reported

II) Sample type, sample size and number of items should be given.

III) The language of the studies should be English or Turkish.

Theses in which Cronbach Alpha reliability coefficient was not given, variables selected as moderator variables within the scope of the research were not reported, and theses published after 2021 were not included in the study. The stages of the meta-analysis process are shown in Figure 1. Within the scope of the research, 78 studies with 81 Cronbach Alpha reliability coefficients that met the inclusion criteria are given in Appendix 1.

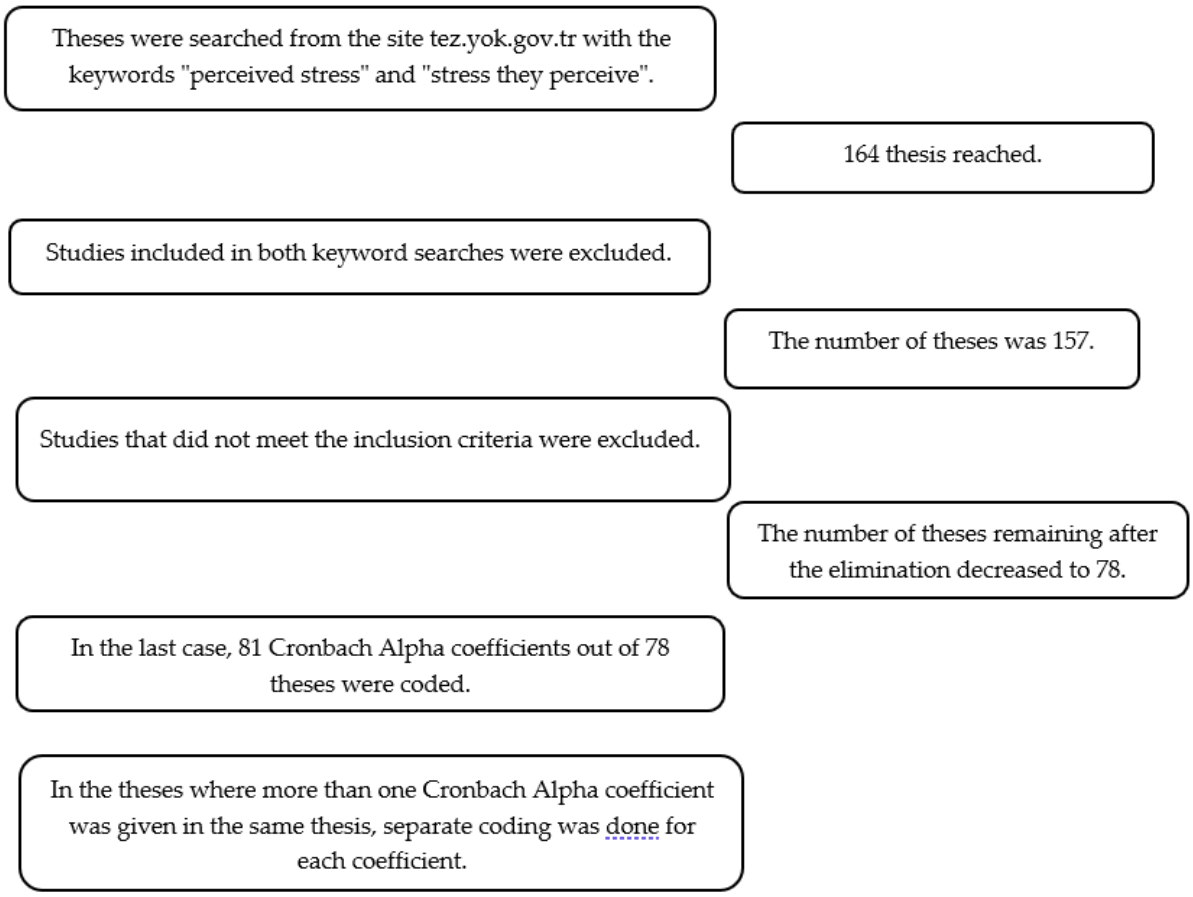


Figure 1. Meta-Analysis stages of studies on perceived stress

The Process of Coding Study Variables

During the evaluation process, data were collected and compared according to the desired criteria in 25 of 78 studies, which is more than 30% of the studies (Şen, 2018). The aim of this process is to examine whether there is agreement between raters. Table 1 and Table 2 display the cross tabulation values and fit statistic values.

Table 1.
Crosstabs between raters

			Rater 2		Total
			0	1	
Rater 1	0	N	14	0	14
		% within Rater 1	100%	0%	100%
		% within Rater 2	93,3%	0%	100%
		% of Total	56%	0%	56%
	1	Count	1	10	11
		% within Rater 1	9,1%	90,9%	100%
		% within Rater 2	6,7%	100%	44%
		% of Total	4%	40%	44%
Total		Count	15	10	25
		% within Rater 1	60%	40%	100%
		%within Rater 2	100%	100%	100%
		% of Total	60%	40%	100%

According to Table 1, rater 1 and rater 2 scored 14 of the 25 studies jointly as "0" and 10 studies jointly as "1". There was disagreement on only one study. The Cohen's Kappa statistic is presented in Table 2.

Table 2.
Cohen's Kappa statistic value

	Value	Asymp. Std. Error	Approx. T	Approx. Sig.
Measure of Agreement Kappa	.92	.08	4.61	.00
N of Valid Cases	25			

Cohen's Kappa Statistic (1960) was used for the percentage of agreement between raters. Accordingly, Cohen's Kappa coefficient was found to be 0.92. This value indicates almost perfect agreement (Landis & Koch, 1977). Following the high agreement values, the variables of interest in all studies were coded. Table 3 displays the descriptive statistics of the studies.

Table 3.
Descriptive statistics of the variables included in the study

Descriptive Variables	Categories	Number of Studies	Number of Cronbach Alpha
Sample Size	<250	39	42
	>250 range <500	29	29
	>500	10	10
Gender	Male/Total	78	81
Number of Items	0-1 range 14	55	56
	10	19	21
	8	4	4
Type of Sample	Student	25	25
	Hospital Staff	18	19
	Individual	26	26
	P-P (Patient or Pregnant)	9	11
Field of Study	Education	14	14
	Health	27	30
	Psychology	26	26
	Business	4	4
	Sport	4	4
	Human Resources	3	3
Publication Year	2005-2010	4	4
	2011-2015	6	6
	2016-2021	68	71
Age of Participants	16-20	10	10
	21-30	27	30
	Over 30	26	26
Type of Thesis	Master	62	63
	Doctoral	12	14
	Medical Specialty	4	4
Total		78	81

Studies were divided into three categories based on sample sizes: small sample (250 or less), medium sample (between 251-500), and large sample (over 500). All of the scales used in the studies belong to Cohen et al. (1983) and forms of the scale consisting of three different numbers of items (14-10-8) were used. The sample of the studies was categorized as students (high school and university), hospital staff

(doctors, nurses, hospital attendants), individuals (teachers, architects, lawyers, parents, patient relatives, etc.), and P-P (patients and pregnant women). A separate category was created because it was thought that the stress experienced by the people in the category called P-P would be different from other individuals, considering their diseases or pregnancy status. This is because pregnancy, which is considered a physiological event, is a source of intense stress for women during the reproductive period (Özkan et al., 2013). Also, Cohen et al. (1983), it was revealed that social tension, depressive disorders, vital events and physical symptoms affect perceived stress. The gender variable was calculated using the overall sample proportion of men. The fields of the studies examined, education, health, psychology, business, sports sciences, and human resources. The publication years of the studies were divided into three categories, with an interval of six years. The reason for separating the years in this way is the wide range of years (17 years) and the fact that the years in which the scale was translated into Turkish were firstly between 2006-2010 and secondly between 2011-2015. The mean ages of the samples in the studies were divided into three categories: 16-20, 21-30 and over 30. The reason for dividing the ages in this way is that the scale is considered as the beginning of high school and early university (16-20), young adulthood (21-30) (Santrock, 2006) and middle age and above (30+) (Boyd & Bee, 2014). Studies were divided into three categories according to the type of thesis: master's, doctoral and medical specialty.

Data Analysis

Considering the argument about the sample-specific nature of reliability, the variation in reliability estimates between studies using the same reliability coefficient can be explained by variation in sample structures or scale forms. Therefore, some sample structures or scale forms may produce higher reliability estimates than others, resulting in differences in reported reliability coefficients between studies. (Vassar & Bradley, 2010). RG is a meta-analytical strategy used to discover variances in reliability coefficients between studies and to identify study variables that may account for such differences. Various techniques have been used for data analysis in RG studies (Vacha-Haase, 1998; Rouse, 2007). In this study, as in many other studies (Aguayo et al., 2011; Allan, 2021; Hongyan et al., 2015; Nicolas et al., 2021; Özdemir et al., 2020; Rouse, 2007, etc.), the Cronbach Alpha coefficients of the studies to be examined were collected. A typical study that measures only a random sample of the universe would include a sampling error in the Alpha coefficient of unknown magnitude and direction (Bonett, 2002). For this reason, instead of using the Alpha coefficient directly in the study, it was converted to other coefficients. The most commonly used of these methods are conversion to Fisher's Z, Hakstian-Whalen and Bonett coefficients. While some of the studies (Beretvas et al., 2002) emphasized the necessity of transforming the Alpha coefficient, it was stated that there was no need to transform the Alpha coefficients in some of the studies (Henson& Thompson, 2002; Thompson& Vacha-Haase, 2000).

For alpha coefficients, Hakstian-Whalen and Bonett transforms are better than Fisher's Z. The Hakstian-Whalen transform normalizes the distribution of the reliability coefficients. However, on a theoretical basis, the Bonett method is a better Alpha coefficient transformation, since the method proposed by Bonett (2002) can normalize the distribution of the Alpha coefficients and fix their variances (Sen, 2021b).

In the study, the Bonett transform was applied to the Alpha coefficients obtained from the studies. In addition, analyzes of the 14-10 and 8-item forms of the PSS were combined. This is because, when the adaptation studies of the scales and their items were examined, the explained variance was found to be similar, and the shorter scale forms were obtained by removing items from the 14-item form. In addition, in studies in which scales with different numbers of items in an analysis were used, it was thought that the number of items in the scales could be moderator variables, and the number of items was included in the moderator analyzes.

The two methods used to obtain the effect size in the meta-analysis are the fixed effects and random effects models. In the fixed-effects model, heterogeneity between studies is low, and variation arises only from the sample participating in the studies. The random effect model, on the other hand, is a model in which the variance between studies is high, and the effect sizes estimated as a result of the meta-analysis are obtained as the average of the differing effect sizes in all studies. In addition, if the model is random at one level and fixed at the other level, this model is called a mixed effects model (Borenstein, 2019). In the random effects model, the variance may also be caused by factors other than the participants in the studies (Sen,2018). Field and Gillett (2010) stated that using the random effects model makes more sense than using the fixed effects model. Since the studies used in the research have different characteristics (sample, field, age, etc.), the random effects model, which assumes that the variance between studies is estimated greater than zero, was used. The heterogeneity of the Alpha coefficient was evaluated by calculating the I2 value as a function of the Q statistic. The Q statistics were applied to test the homogeneity assumption between the alpha values. While the I2 statistic is a possible measure of the amount of heterogeneity according to Higgins and Thompson (2002), Borenstein (2019) stated that the I2 statistic does not tell the researcher how much the effect size varies but provides information about the relationship between two distributions. It can be said that I2 values of approximately 25%, 50% and 75% reflect low, medium and large heterogeneity, respectively (Huedo-Medina et al., 2006).

Before interpreting the obtained effect size conclusions, Bonett-transformed values were converted back to Alpha coefficients. The impact of moderator variables on the variability of the reliability estimates was assessed using analog ANOVA and meta-regression. The moderator variables—sample size (≤250, 251-500, >500), number of items (14, 10, 8), type of sample (students, hospital staff, individuals, P-P), field of study (education, health, psychology, business, sports, human resources), publication year (2005-2010, 2011-2015, 2016-2021), age (16-20, 21-30, >30), and type of thesis (master's, doctoral, medical specialty)—were analyzed using analog ANOVA, while the gender ratio (number of men/sample size) was analyzed using meta-regression. Finally, analyzes were made using fail-safe N method, Orwin's fail-safe N (Orwin, 1983), Begg and Mazumdar's rank correlation test (Begg & Mazumdar, 1994), Duval and Tweedie's Trim and Fill method (Duval & Tweedie, 2000), Egger's regression test (Egger et al., 1997), and funnel plot methods to examine publication bias.

Fail-Safe N aims to provide assurance that the results are not entirely an artifact of publication bias. The fact that the p-value for Fail-Safe N is smaller than the alpha value ($p<0.001$) indicates that the study is a strong study with low reliability (Borenstein, 2019; Eser & Doğan, 2023). Begg and Mazumdar suggested calculating the rank correlation between precision and effect size. A statistically significant correlation indicates that the average effect size is larger in small studies. The Trim and Fill method presumes that area studies are missing due to publication bias. It creates these studies, adds them to the analysis and runs the analysis using the original and assigned studies to obtain an adjusted mean. If the p-value obtained from Egger's regression test is lower than the alpha level ($p<.05$), it indicates the presence of publication bias. (Borenstein, 2019, Şen & Yıldırım, 2020). CMA package program was used for statistical analyses.

Results

The descriptive statistics of the different item numbers in the scales for the 81 Alpha coefficients of the 78 theses examined within research scope are given in Table 4. Values for each scale were calculated separately, and then overall values were obtained for all studies.

Table 4.
Descriptive statistics of the theses examined within the research scope

	Number of Cronbach' α	Mean Cronbach' α	Lower 95% CI	Upper 95% CI	Min	Max
PSS-14	56	.80	.78	.82	.60	.97
PSS-10	21	.83	.81	.84	.76	.88
PSS-8	4	.78	.71	.85	.66	.86
Total	81	81	.79	.82	.60	.97

According to Table 4, PSS-14 was included in 56 studies and had Cronbach alpha values between .60-.97. These values are the lower and upper limits in all studies. PSS-10 was used in 21 studies and values between .76-.88 were obtained. PSS-10 had the highest lower and upper limit values. PSS-8 was used in only 4 studies and had the lowest mean alpha value and the lowest lower limit value. The stem-and-leaf plot of the Alpha coefficients obtained from the studies is as in Figure 2.

```
 Frequency   Stem & Leaf

    2,00 Extremes    (=<,61)
    4,00      6 .  6789
    9,00      7 .  112223344
   19,00      7 .  5555667888888899999
   22,00      8 .  0011112222334444444444
   19,00      8 .  5555666677777777899
    5,00      9 .  00111
    1,00      9 .  7

 Stem Width:      ,10
Each Leaf:       1 case
```

Figure 2. Stem-and-leaf plot of studies

According to the stem-and-leaf plot, the distribution appears to be close to normal (slight negative skew (left-skewed)) and the density is concentrated in the .80-.85 range, with 22 studies. Descriptive statistics of the theses including mean, standard deviation, skewness and kurtosis values are given in Table 5.

Table 5.
Descriptive Statistics of Studies

	N	Mean	SD	Skewness		Kurtosis	
PSS	81	.81	.07	Statistic	Std. Error	Statistic	Std. Error
				-.61	.28	.46	.53

According to Table 5, the mean alpha coefficient of 81 studies is .81 and the standard deviation is .07. An analysis of the skewness and kurtosis values reveals that the distribution of Alpha coefficients is slightly left-skewed (negatively skewed). Table 6 presents the descriptive statistics for the general and moderator variables converted into Bonett coefficients, as well as the values converted back into Alpha coefficients. It also shows the lower and upper limits, and the minimum and maximum values of the Alpha coefficients for the studies analyzed within the scope of the research.

Table 6.
Reliability estimates of PSS across studies for different moderator variables

			%95 Confidence Interval			
Moderator Variables	Mean Effect Size	Bonett	Lower Bound	Upper Bound	Minimum	Maximum
Sample Size						
<250	.83	-1.77	.80	.85	.68	.97
250-500	.81	-1.65	.78	.83	.60	.87
>500	.81	-1.66	.76	.85	.66	.87
Number of Items						
14	.82	-1.69	.79	.84	.60	.97
10	.83	-1.78	.82	.85	.76	.87
8	.79	-1.56	.68	.86	.66	.86
Type of Sample						
Student	.83	-1.76	.80	.85	.67	.91
Hospital Staff	.82	-1.71	.79	.85	.66	.89
Individual	.82	-1.72	.78	.85	.60	.97
P-P	.78	-1.52	.76	.80	.72	.87
Field of Study						
Education	.82	-1.69	.78	.85	.67	.87
Health	.81	-1.67	.79	.84	.66	.91
Psychology	.84	-1.82	.82	.86	.68	.91
Business	.86	-1.94	.63	.95	.71	.97
Sport	.70	-1.21	.65	.75	.60	.74
Human Resources	.79	-1.56	.67	.87	.68	.86
Publication Year						
2005-2010	.83	-1.78	.82	.85	.81	.84
2011-2015	.83	-1.79	.80	.86	.76	.87
2016-2021	.82	-1.70	.80	.83	.60	.97
Age of Participants						
16-20	.81	-1.68	.77	.85	.67	.87
21-30	.82	-1.73	.80	.84	.60	.91
over 30	.81	-1.67	.80	.84	.69	.91
Type of Thesis						
Master	.82	-1.72	.71	.89	.61	.97
Doctoral	.79	-1.58	.74	.84	.60	.84
Medical Specialty	.83	-1.77	.67	.91	.69	.88
Total	.82	-1.71	.80	.83	.60	.97

In Table 6, the statistical values of the moderator variables and the mean effect size coefficient for the total PSS scores are presented in the bottom line. While the mean effect size coefficient of the PSS is .82, the lower bound is .80 and the upper bound is .83 in the 95% confidence interval. In parallel, the scores of the studies ranged between .60 and .97 in reliability. The mean effect size values of the studies divided into three categories according to the sample size are between .81 and .83. The mean effect size values of the item numbers of the scales used in the studies are between .79 and .83. The mean effect size values of types of samples of the studies are between .78 and .83. The mean effect size values of the fields of the studies are between .70 and .84. The mean effect size values of the studies divided into three categories according to the publication years are between .82 and .83. The ages of the participants in the studies were divided into three categories and the mean effect size values of these categories were between .81 and .82. According to the type of studies, the mean effect size values are between .79 and .83.

The heterogeneity of the effect size values of the reliability coefficients within the research scope was analyzed through the Q and I2 statistics and the forest plot. The Q statistic for the effect sizes $Q(80) = 1543.26$, $p < .01$ was found to be statistically significant. The I2 value of the effect sizes was found to

be 94.82. If this value is close to 0, it indicates low heterogeneity, and if it is close to 1, it indicates high heterogeneity. (Şen &Yıldırım, 2020). In addition, the forest plot of the studies examined within the scope of the research is given in Figure 3.

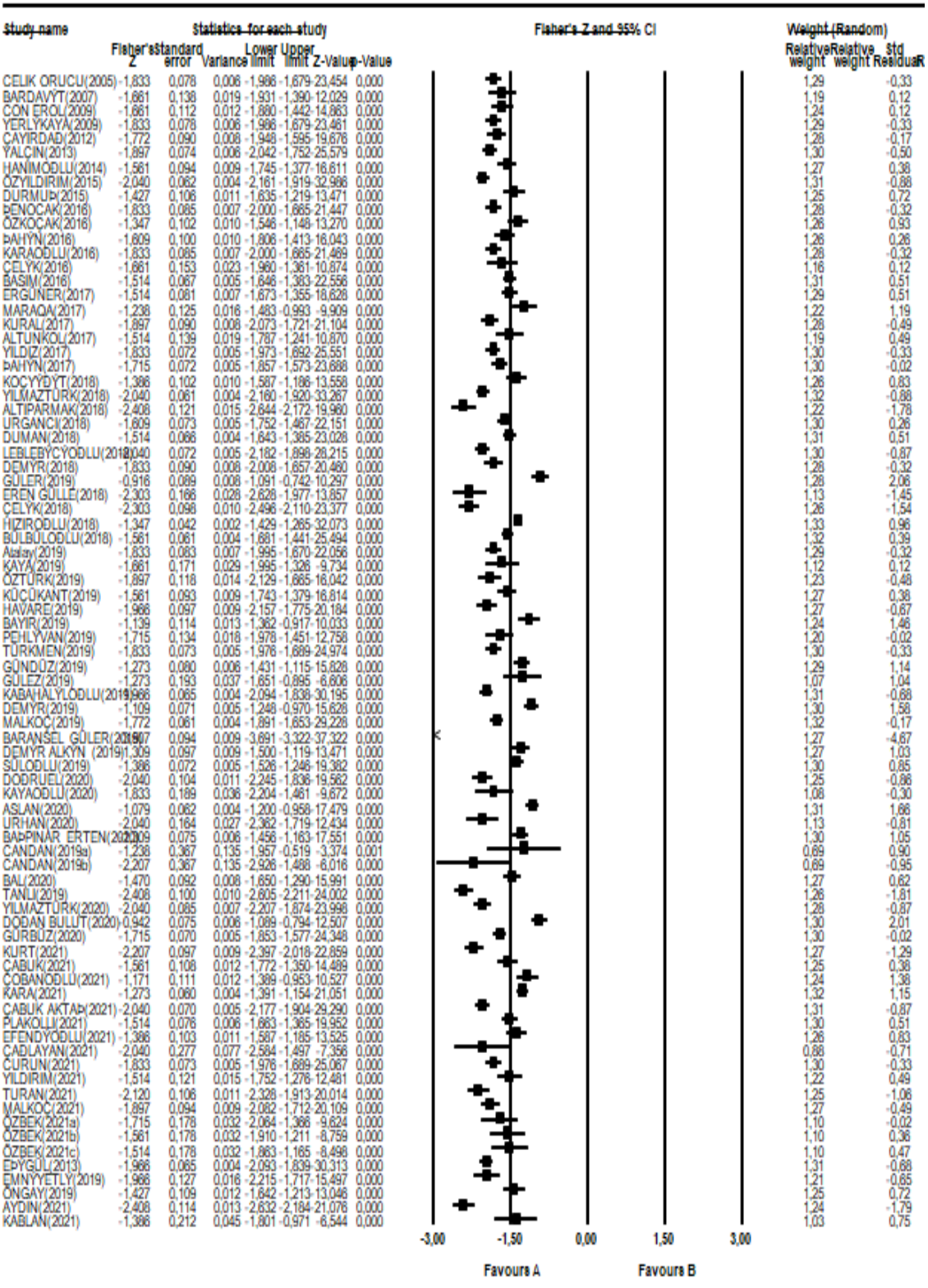


Figure 3. Forest plot of studies

Given that the results of the analysis indicate that there is heterogeneity across studies that is worth examining, the sources of heterogeneity need to be explained. For this reason, Analog ANOVA and meta-regression analyses were performed on moderator variables to determine possible causes of heterogeneity between effect sizes. Analogue ANOVA tests the homogeneity of effect sizes for subcategories of a categorical variable and the differences between categories (Lipsey & Wilson, 2001). When analyzing the results, a significance level of $p < .05$ is used. If $p < .05$, the moderator variable is considered a source of heterogeneity; otherwise, it is not. Table 7 displays the analysis results for the sample size moderator variable.

Table 7.
Analog ANOVA results of the sample size variable

Model	Sample Size	Q	df(Q)	p-value
Fixed Effect	<250	733.86	41	.00
	250-500	437.93	28	.00
	>500	311.69	9	.00
	Total within	1483.47	78	.00
	Total between	59.95	2	.00
	Overall	1543.43	80	.00
Mixed Effect	Total between	1.61	2	.45

When Table 7 is examined, it is observed that the p-value is greater than .05 according to the mixed effects model, indicating that the sample size moderator variable is not a source of heterogeneity. Table 8 displays the results of the Analog ANOVA, where the heterogeneity of the number of items in the scales is examined as a moderator variable.

Table 8.
Analog ANOVA results of the number of items variable

Model Type	Number of Items	Q	df(Q)	p-value
Fixed Effect	14	1284.66	55	.00
	10	111.21	20	.00
	8	50.85	3	.00
	Total within	1446.72	78	.00
	Total between	96.71	2	.00
	Overall	1543.43	80	.00
Mixed Effect	Total between	1.92	2	.38

According to the results in Table 8, it was determined that the p-value was greater than .05 in the mixed effects model and the moderator variable of the item numbers of the scales was not a source of heterogeneity. Table 9 displays the results of Analog ANOVA in which the heterogeneity of types of samples included in the studies was examined as a moderator variable.

Table 9.
Analog ANOVA results of types of samples variable

Model Type	Types of Samples	Q	df(Q)	p-value
Fixed Effect	Individual	809.11	25	.00
	Hospital Staff	250.46	18	.00
	P-P	16.76	10	.08
	Student	441.14	24	.00
	Total within	1517.46	77	.00
	Total between	25.96	3	.00
	Overall	1543.43	80	.00
Mixed Effect	Total between	9.71	3	.02

When Table 9 is examined, it is observed that the p-value is less than .05 according to the mixed effects model and the types of samples moderator variable is one of the sources of heterogeneity. The results of the Analog ANOVA, where the heterogeneity of the fields of the studies was examined as the moderator variable, are presented in Table 10.

Table 10.
Analog ANOVA results of the fields of the studies variable

Model Type	Field of Study	Q	df(Q)	p-value
Fixed Effect	Education	244.69	13	.00
	Human Resources	30.74	2	.00
	Business	379.09	3	.00
	Psychology	265.38	25	.00
	Health	349.18	29	.00
	Sport	19.17	3	.00
	Total within	1288.25	75	.00
	Total between	255.18	5	.00
	Overall	1543.43	80	.00
Mixed Effect	Total between	37.89	5	.00

According to the results in Table 10, the p-value was found to be less than .05 in the mixed effects model, indicating that the study fields moderator variable was a source of heterogeneity. In Table 11, the results of the analysis of the moderator variable of the publication years of the studies are given.

Table 11.
Analog ANOVA results of publication year variable

Model Type	Publication Year	Q	df(Q)	p-value
Fixed Effect	2005-2010	2.79	3	.43
	2011-2015	39.20	5	.00
	2016-2021	1460.22	70	.00
	Total within	1502.21	78	.00
	Total between	41.21	2	.00
	Overall	1543.43	80	.00
Mixed Effect	Total between	1.74	2	.42

According to the results in Table 11, the p-value was found to be greater than .05 in the mixed effects model, indicating that the publication years of the studies were not a source of heterogeneity as a moderator variable. While the gender of the participants in the studies was used as a moderator variable, the results obtained by taking the ratio of male participants to all participants were analyzed by meta-regression analysis. Table 12 displays the results of the meta-regression analysis.

Table 12.
Meta-Regression analysis results for the gender moderator variable based on the mixed effects model

	Coefficient	Std. Error	95% Lower	95% Upper	Z-value	p-value
Intercept	-1.68	.09	-1.85	-1.51	-19.3	.00
Gender(male%)	-0.05	.20	-.43	.34	-.23	.82

When the results in Table 12 are examined, it is observed that the p-value is greater than .05 in the mixed effects model, indicating that the gender of the participants is not a source of heterogeneity as a moderator variable. Table 13 presents the results of Analog ANOVA, in which the heterogeneity of the age of the participants of the studies was examined as a moderator variable.

Table 13.
Analog ANOVA results for the age variable

Model Type	Age of Participants	Q	df(Q)	p-value
Fixed Effect	16-20	147.98	10	.00
	21-30	418.17	29	.00
	Over 30	347.03	25	.00
	Total within	913.18	63	.00
	Total between	.82	2	.66
	Overall	914.01	65	.00
Mixed Effect	Total between	.47	2	.79

When the results in Table 13 are examined, it is seen that the p-value is greater than .05 in the mixed effects model and the moderator variable of the participants' age is not a source of heterogeneity. Table 14 presents the results of Analog ANOVA, in which the heterogeneity of the types of theses is examined as a moderator variable.

Table 14.
Analog ANOVA results of types of theses variable

Model Type	Type of Thesis	Q	df(Q)	p-value
Fixed Effect	Master	1295.14	62	.00
	Doctoral	108.26	13	.00
	Medical Specialty	51.03	3	.00
	Total within	1455.43	78	.00
	Total between	98.80	2	.00
	Overall	1554.23	80	.00
Mixed Effect	Total between	1.91	2	.52

When the results in Table 14 are examined, it is observed that the p-value is greater than 0.05 in the mixed effects model and the moderator variable of the theses is not a source of heterogeneity. Within the scope of the research, the generalization of reliability was conducted using theses from Türkiye with the PSS developed by Cohen et al. (1983). Since the majority of published studies have high or significant effect sizes, conducting a meta-analysis solely on these studies may lead to publication bias. Therefore, publication bias analysis was performed. Funnel plot was examined for publication bias, then Egger's regression test, Duval and Tweedie's Trim and Fill method, Begg and Mazumdar's rank correlation test, Orwin's safe N and fail-safe N methods and p curve analysis were used. In the fail-safe N method, assuming that the main effect of the additional studies is zero, the number of studies needed to render the p-value insignificant is calculated, and this number is referred to as the safe N. If only a few studies are required, there may be concern that the effect is actually zero. (Borenstein et al., 2013). With this method, N number at $p < .01$ level was calculated as 13740. In Orwin's safe N method, the N value is 9114. The funnel plot asymmetry is shown in Figure 4.

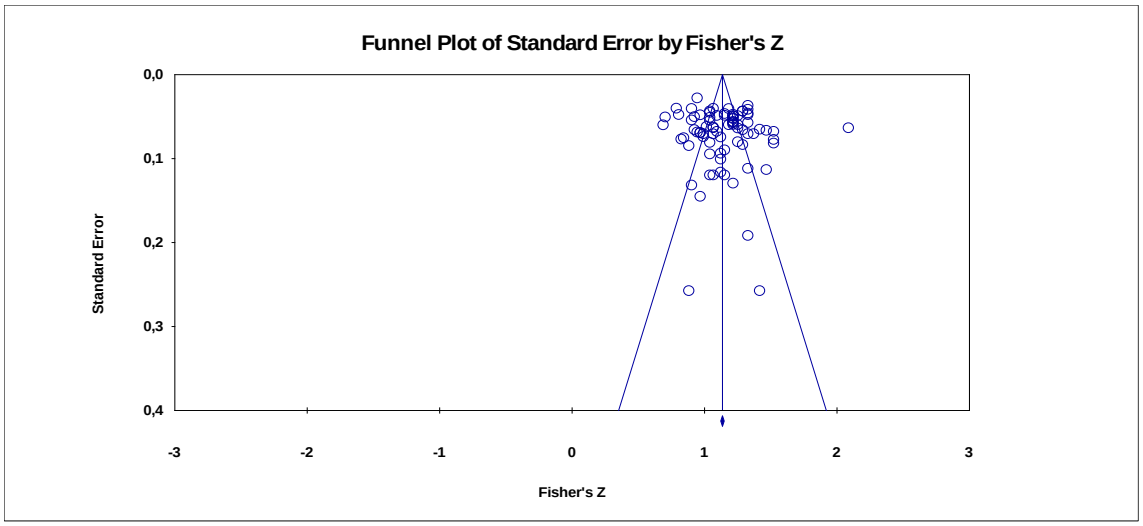


Figure 4. Funnel plot of effect sizes

If there is no publication bias, studies will be distributed symmetrically according to the mean effect size as the sampling error is random (Borenstein et al., 2013). Although the studies appear to be symmetrically distributed on both sides of the effect size, the interpretation is not entirely objective. According to the results of Egger's regression test, the regression cut-off point (intercept) was not significant (cut-off point = 1.230, p = 0.31). This indicates that the regression constant does not deviate significantly from zero. Begg and Mazumdar's rank correlation test also contributed to other results. Accordingly, Kendall's tau value is not significant (Kendall's tau = 0.004, p = 0.96). Finally, the Duval and Tweedie Trim and Fill test found no difference between the observed effect size and the adjusted effect size created to correct for publication bias. As a result of the general symmetrical distribution of the studies performed on both sides of the overall effect size, the difference was found to be zero. According to the results obtained from the analysis of all publication bias, it can be said that there is no evidence of publication bias in the studies. The p-curve represents the distribution of statistically significant p-values ($p < .05$) across a set of studies (Simonsohn et al. 2014). The p-curve visual is given in Figure 5.

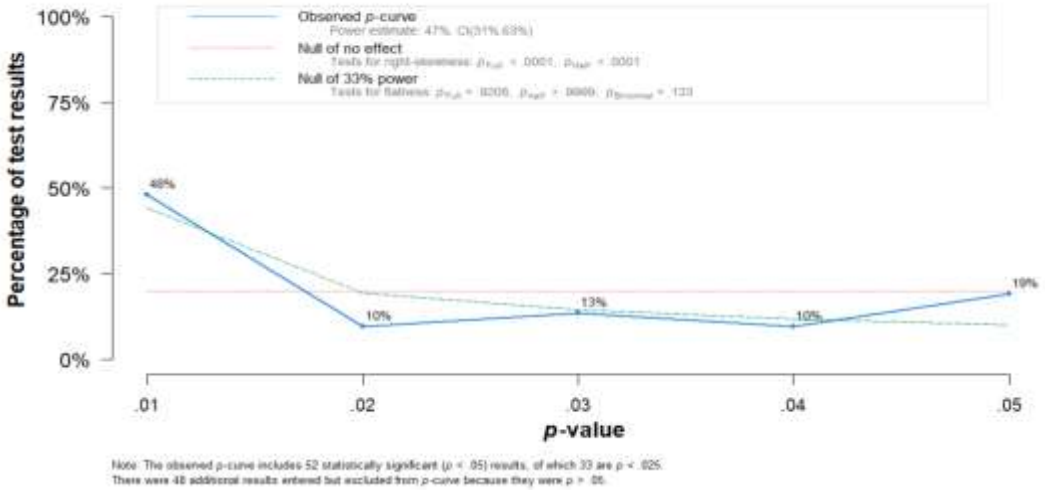


Figure 5. p-curve plot

The p-curve output was interpreted as another analysis to examine publication bias. The p-curve given in Figure 5 contains 52 studies at the $p < 0.05$ significance level, 33 of which have a $p < 0.025$ significance level. Since the p-values of 48 studies was greater than 0.05, these studies were not evaluated. The blue line represents the p-curve with a 48% power estimate. The sharp drop in the p-curve and the concentration of results at lower p-values, together with the statistical tests, indicate that the effects seen in the studies are real and not due to randomness or publication bias. However, the drop in the figure is not sharp and gives an indication that publication bias may be present.

Discussion

In this study, a meta-analytic reliability generalization analysis was conducted. Additionally, it was investigated whether the Cronbach's Alpha coefficient was influenced by sample size, number of items in the scale, sample type, field of study, year of publication, gender, age and type of thesis. The results of this study showed that the average Cronbach's Alpha coefficients obtained from the PSS were at an acceptable level. The fact that these coefficients are generally high can be considered as an indicator of the usability of the scale by both practitioners and researchers. With RG, the average reliability coefficient of the studies can be calculated and the moderator variables that will reveal the variability between the studies can be determined. In this context, RG studies of Maslach Burnout Inventory (MBI) and Beck Depression Inventory-II (BDI), which are similar to PSS, were examined and reliability coefficients and moderator variables that may cause variability in these coefficients were compared. RG studies of Maslach Burnout Inventory were conducted by Alenezi (2023), Wheeler et al. (2011) and Aguayo et al. (2011). In these studies, the mean reliabilities of the sub-dimensions of the scales (Emotional Exhaustion, Depersonalization and low Personal Accomplishment) were found to be .83, .78 and .77 by Alenezi (2023), .87, .71 and .76 by Wheeler et al. (2011) and .88, .71 and .78 by Aguayo et al. (2011) respectively. The reliability coefficients of the original scale are .89, .77 and .74. The RG study of the BDI-II was conducted by Eser and Aksu (2021). In this study, the average reliability coefficient of the studies was found to be .90. The reliability coefficient of the original scale is .92. The average reliability coefficient obtained from three different forms of PSS in this study is 0.81. The reliability coefficient of the original scale is .84. Accordingly, the average reliability coefficients of the PSS and the other two scales are close to the reliability coefficients of the original scales. However, the high reliability coefficients of all three forms of the PSS provide an advantage over the others. It was found that there was heterogeneity in the studies included in the research and as a result of the investigation of the sources of this heterogeneity, it was concluded that two of the eight moderator variables (type of sample and field of study) were the source of heterogeneity.

The structure that PSS tries to measure is stress, and it has been stated in the literature that patients and pregnant women in the P-P category in the type of sample moderator variable experience more stress than other individuals. Therefore, careful consideration is required when selecting samples for studies using the PSS. Özdemir et al. (2020) found the type of sample as a source of heterogeneity in their reliability generalization study for the short and long forms of the Oxford Happiness Scale. The type of sample was found to be a source of heterogeneity in Alzahrani's (2016) thesis in which he conducted a generalization study of reliability to the Brief Symptom Inventory-18 scale. In the reliability generalization studies of Alenezi (2023) and Aguayo et al. (2011) on the Maslach Burnout Inventory (MBI), the sample type was again identified as a source of heterogeneity. Field of study is another variable that may cause heterogeneity when using the PSS. Lower effect size values were obtained than other fields in sports sciences and human resources management in this research. Similarly, Özdemir et al. (2020) and Wheeler et al. (2011) identified field of study as a source of heterogeneity.

In addition, it was concluded that the variables of sample size, gender, age, publication year, type of thesis and number of items in the scale were not sources of heterogeneity in the scope of PSS. Eser and Aksu (2021) found that none of the moderator variables (gender, publication year, type of sample and language) had an explanatory role in the reliability generalization study of Beck Depression Inventory II. In the reliability generalization study by Vassar and Crosby (2008) on the UCLA Loneliness Scale, study type was identified as a source of heterogeneity. Özdemir et al. (2020) found the moderator variable as a source of heterogeneity. In the study of Alzahrani (2016), the gender variable was found

as a source of heterogeneity. In the reliability generalization study of Hongyan et al.'s (2015) 44-item Big Five Inventory, gender, age, sample size and nationality of the participants were found to be sources of heterogeneity. In the study of Aguayo et al. (2011), the country where the study was conducted, the age of the participants, the language, and the type and version of the inventory were found to be sources of heterogeneity. Sample size, age and geographical region were found to be sources of heterogeneity in Allan's (2021) thesis in which she examined the validity and reliability generalization of the Adult Characterological Measurement of Resilience. In addition to the moderator variables examined, the moderator variables varied depending on the scale or inventory examined in different studies. For example, in Esparza-Reig et al. (2021) reliability generalization study of the South Oaks Gambling Screen, the continent where the study was conducted, the data collection method (face to face-others) and clinical conditions moderator variables were found to be sources of heterogeneity. Vassar and Crosby's (2008) study found that the moderator variables of adolescence and social network disconnection are sources of heterogeneity. Vassar and Bradly (2010) found that the moderator variables of language and adolescence are the source of heterogeneity as a result of the reliability generalization they conducted in their research on the Life Orientation Test.

Since the number of items in the scale was not found to be a source of heterogeneity, researchers using the PSS can choose any form (14, 10, or 8 items). A similar conclusion applies to the sample size, gender, and age variables. However, the results in this study were obtained specific to the conditions in the theses examined. Therefore, the results should be considered within the scope of the examined studies, and researchers using the PSS should calculate and interpret reliability based on their own data. In addition, PSS is a scale that is used quite frequently in studies, and for this reason, it is very difficult to address all the studies that include PSS for the RG study. To overcome this situation, only theses were used within the scope of the study. To improve the representativeness of the dataset in an RG study, the scale selected for reliability analysis should have a sufficient number of studies—neither too many nor too few—to allow for generalizations. In addition, in the selection phase, papers can also be included in the scope of the studies to be analyzed. In conclusion, this study of the theses using the PSS demonstrates that the PSS is a reliable scale that can be generalized to different contexts. In the light of these findings, PSS can be used effectively regardless of age range, gender, thesis type, sample size (provided that it is not too small to affect the reliability coefficient) and number of items (8-10-14).

Within the scope of the study, in addition to master's and doctoral dissertations, only medical specialty studies (four) were used within the scope of gray literature sources in the thesis center. The reason for this can be said to be the high number of studies using PSS as mentioned before. In this study eight moderator variables that could be a source of heterogeneity for the theses made using PSS (improved by Cohen et al. (1983)) were examined and the Cronbach Alpha coefficients were transformed with the Bonett method and included in the analysis. In sample selection, it should be realized that the perceived stress levels of patients and pregnant women are higher than other individuals, and research should be conducted by knowing that perceived stress differs in the fields of sports sciences and human resources compared to other fields. Because these variables may change the reliability coefficient of PSS in the studies in which they are included. Unlike PSS, prenatal perceived stress scale can be used for clinical situations such as the effect of prenatal stress perceived by pregnant women on the mother's attachment to the fetus or marital adjustment (Çalışkan Altıntaş, 2024; Yüksel, 2024). However, PSS can be used in studies conducted for individuals with obsessive disorders, bipolar disorders, etc. (Acar, 2024; Toprak, 2023). Nevertheless, in recent years, studies conducted on different samples such as law enforcement officers, refugees and people with post-covid trauma disorders can be examined and the type of sample can be expanded. In studies conducted for patients, the type of disease can be considered as a source of heterogeneity. Because differences in the recovery of diseases may affect perceived stress. In addition, in future studies in this field, variables such as the language in which the scales are applied, ethnicity if different, marital status according to the group to which the scales are applied and research design can be examined by identifying them as different sources of variability. Lastly in future studies, reliability coefficients can be analyzed using different transformation methods, such as the Hakstian-Whalen method.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Author Contribution: Ömer DOĞAN: conceptualization, investigation, methodology, data curation, writing - review & editing, visualization. Selahattin GELBAL: conceptualization, supervision, formal analysis, review & editing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as this study uses data shared with the public.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Acar, O. (2024). *The relationship between emotional eating, perceived stress and perceived social support in people with and without a diagnosis of bipolar disorder.* [Unpublished Master Thesis]. İstanbul Arel University.

Agbo, A. A. (2010). Cronbach's alpha: review of limitations and associated recommendations, *Journal of Psychology in Africa, 20*(2), 233-239. http://dx.doi.org/10.1080/14330237.2010.10820371

Aguayo, R., Vargas, C., & Lozano, L. M., (2011). A meta-analytic reliability generalization study of the maslach burnout inventory. *International Journal of Clinical and Health Psychology, 11*(2), 343-361.

Alenezi, S. S. (2023). The reliability of Maslach burnout inventory in Arab studies: Reliability generalization meta-analytic study. *International Journal for Research in Education, 47*(5),46-75. http://doi.org/10.36771/ijre.47.5.23-pp46-75

Allan, T. A. (2021). *Resiliency: a systematic review of adult characterological measures of resilience and reliability and validity generalization studies of the brief resilience scale*. [Unpublished Doctoral Dissertations]. University of Ottawa Social Sciences.

Alzahrani, D. T. (2016). *A reliability generalization study of the brief symptom inventory-18*. [Unpublished Master Thesis]. University of Denver.

Andreou, E., Alexopoulos, E. C., Lionis, C., Varvogli, L., Gnardellis, C., Chrousos, G. P. & Darviri, C. (2011). Perceived stress scale: reliability and validity study in greece. *International Journal of Environmental Research and Public Health*, *8*(8):3287-3298. https://doi.org/10.3390/ijerph8083287

Ataman Temizel, E. & Dağ, İ. (2014). Relationships between stressful life events, cognitive emotion regulation strategies, depressive symptoms and anxiety level. *Journal of Clinical Psychiatry*, *17*(1).

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088-1101. doi: 10.2307/2533446

Beretvas, S. N., Meyers, J. L., & Leite,W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 62*(4), 570–589.

Beretvas, S. N., Suizzo, M., Durham, J. & Yarnell, L. M. (2008). A reliability generalization study of scores on rotter's and nowicki-strickland's locus of control scales. *Educational and Psychological Measurement, 68*(1), 97-119.

Bilge A., Öğce, F., Genç, R. E., & Oran, N. T. (2009). Psychometric relevance of the Turkish version of the Perceived Stress Scale (PSS). *Ege University School of Nursing Journal, 2*(25), 61-72

Birol, L., & Akdemir, N. (2003). *Internal Medicine and Nursing Care* 1. Edition, Vehbi Koç Foundation Publications.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics 27*(4): 335-340.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2013). *Introduction to meta-analysis*. UK: John Wiley & Sons.

Boyd, D., & Bee, H. (2014). *Lifespan Development (7th ed.).* New York, NY: Pearson International.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. & Demirel, F. (2019). *Scientific research methods in education.* Pegem Academy.

Camci, G. B. (2021). *Determination of the relationship between job stress and burnout levels of nurses and their occupational and life satisfaction levels*. [Unpublished Master Thesis]. Atatürk University.

Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement 20*(1), 37-46

Cohen S. (1994), Perceived stress scale, Retrieved from https://www.mindgarden.com/documents/PerceivedStressScale.pdf on 18.10.23.

Cohen, S., Kamarck, T. & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(1), 386-396.

Cohen, S., Kessler, R.C. & Gordon, L.U. (1997). Strategies for measuring stress in studies of psychiatric and physical disorders. S. Cohen, R.C. Kessler & L.U. Gordon (Ed.), *Measuring stress: A guide for health and social scientists* 122–148. New York: Oxford University Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, & Winston.

Cronbach, L. J. (1947). Test "reliability": its meaning and determination. *Psychometrica 12*, 1-16.

Çalışkan Altıntaş, D. (2024). *The relationship between marital adjustment and perceived prenatal stress in pregnant women,* [Unpublished Master Thesis], Tokat Gaziosmanpaşa University.

Çelik-Örücü, M., & Demir, A. (2009). Psychometric evaluation of perceived stressscale for Turkish university students. *Stress and Health, 25*(1), 103-109.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463. doi: 10.1111/j.0006-341x.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj, 315*(7109), 629-634. doi: 10.1136/bmj.315.7109.629

Erci B. (2006). Reliability and validity of the Turkish version of perceived stress scale. *A.Ü. Journal of the School of Nursing, 9*(1):58-64.

Eser, M. T. & Aksu, G. (2021). Beck depression inventory: a study for meta-analytical reliability generalization. *Pegem Journal of Education and Instruction, 11*(3),88-101

Eskin, M., Harlak H., Demirkıran, F. & Dereboy, Ç. (2013). Adaptation of the perceived stress scale to Turkish: reliability and validity analysis. *New Symposium Journal 51* (3): 132-140.

Esparza- Reig, J., Guillen-Riquelme, A., Marti-Vilar, M. & Gonzalez-Sala, F. (2021). A reliability generalization meta-analysis of the south oaks gambling screen (sogs). *Psicothema 33*(3). 490-499.

Field, A. & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694.

Gallagher, M. W., Zvolensky, M. J., Long, L. J., Rogers, A. H. & Garey, L. (2020). The impact of covid-19 experiences and associated stress on anxiety, depression, and functional impairment in American adults. *Cognitive Therapy and Research, 44*(6). https://doi.org/ 10.1007/s10608-020-10143-y.

Graves, B.S., Hall. M.E., Dias-Karch, C., Haischer, M.H. & Apter, C. (2021). Gender differences in perceived stress and coping among college students. *Plos One*. *16*(8): e0255634. https://doi.org/10.1371/journal.pone.0255634

Haertel, E. (2006). Reliability. Brennan, R. L. (2006). *Educational measurement* (65-110). 4th ed. Westport (Conn.): Praeger Publishers.

Havare, G. (2019). *The effect of perceived stress level on job satisfaction: a study on nurses working in a public hospital*. [Unpublished Master Thesis]. Bahçeşehir University.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*(2), 113–127

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. doi: 10.1002/sim.1186

Hongyan, L., Jianping, X., Jiyue, C. & Yexin, F. (2015). A reliability meta-analysis for 44 items big five inventory: based on the reliability generalization methodology. *Advances in Psychological Science. 23*(5), 755-765.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*(2), 193-206. doi: 10.1037/1082-989X.11.2.193

Hussey, I., Alsalti, T., Bosco, F., Elson, M., & Arslan, R. C. (2023). An aberrant abundance of Cronbach's alpha values at .70. Retrived October 05, 2024, from https://osf.io/preprints/psyarxiv/dm8xn, https://doi.org/10.31234/osf.io/dm8xn

Katsarou, A. L., Triposkiadis, F. & Panagiotakos, D. (2013). Perceived stress and vascular disease: where are we now? *Angiology, 64*(7), 529-534. doi:10.1177/0003319712458963

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1), 159-174.

Lazarus, R. S. (1990). Theory-based stress measurement. *Psychological Inquiry, 1*(1), 3–13. https://doi.org/10.1207/s15327965pli0101_1

Lazarus, R.S. (1993). From psychological stress to the emotions: a history of changing outlooks. *Annual Reviews Psychology, 44*, 1- 21.

Lazarus, R. S. & Launier, R. (1978). Stress-related transactions between person and environment. *Perspectives in Interactional Psychology*, 287–327. doi:10.1007/978-1-4613-3997-7_12

Lee, J., Joo, E. & Choi, K. (2012). Perceived stress scale and self-esteem mediate the effects of work-related stress on depression. *Stress & Health 29*(1) 75-81. **https://doi.org/10.1002/smi.2428**

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of Clinical Epidemiology, 62*(10), 1-34. doi: 10.1016/j.jclinepi.2009.06.006

Lipsey, M. W. & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

Nicolas, N. L., Aparicio, M. R., Ibanez, C. L.& Meca, J. S. (2021). A reliability generalization meta-analysis of the dimensional obsessive-compulsive scale. *Psicothema, 33*(3),481-489.

Orwin, R.G. (1983). A fail-safe N for effecet size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157-159.

Osmanovic-Thunström, A., Mossello, E., Akerstedt, T., Fratiglioni, L. & Wang, H. (2015). Do levels of perceived stress increase with increasing age after age 65? A population-based study, *Age and Ageing, 44*(5), 828–834. https://doi.org/10.1093/ageing/afv078

Özdemir, V., Yıldırım, Y., & Tan, Ş. (2020). A meta-analytic reliability generalization study of the Oxford Happiness Scale in Turkish sample. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 374-404. doi: 10.21031/epod.766266.

Özkan, S., Sakal, F. N., Avcı, E. Civil, E. F. & Tunca, M. Z. (2013). Preference of women's birth method and related factors. *Turkish Journal of Public Health, 11*(2), 59–71.

Pedrozo-Pupo, J. C., Pedrozo-Cortes, M. J., Campo-Arias, A. (2020). Perceived stress associated with covid-19 epidemic in Colombia: an online survey. *Cad Saúde Pública, 36* (5), https://doi.org/10.1590/0102-311x00090520.

Qiu, J., Shen, B., Zhao, M., Wang, Z., Xie, B. & Xu, Y. (2020). A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *General Psychiatry, 33*(2), https://doi.org/ 10.1136/gpsych-2020-100213.

Ramírez, M. T. G., & Hernández, R. L. (2007). Factor structure of the perceived stress scale (PSS) in a sample from Mexico. *The Spanish Journal of Psychology, 10*(1), 199–206. doi:10.1017/S1138741600006466

Razurel, C., Kaiser, B., Dupuis, M., Antonietti, J., Citherlet, C. Epiney, M. & Sellenet, C. (2014). Validation of the antenatal perceived stress inventory. *Journal of Health Psychology*, *19*(4), 471-481. doi:10.1177/1359105312473785

Rouse, S. V. (2007). Using Reliability Generalization methods to explore measurement error: An illustration using the MMPI–2 PSY–5 scales. *Journal of Personality Assessment, 88*, 264–275.

Santrock, J. W. (2012). *Life-span developmental Psychology*. New York: McGraw Hill Companies, Inc.

Sheu, S., Lin, H. S., & Hwang, S. L. (2002). Perceived stress and physio-psycho-social status of nursing students during their initial period of clinical practice: the effect of coping behaviors. *International journal of nursing studies*, *39*(2), 165–175. https://doi.org/10.1016/s0020-7489(01)00016-5

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve and effect size: correcting for publication bias using only significant results. *Perspectives on psychological science: a journal of the Association for Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Singh, S., Pandey, N. M., Datta, M. & Batra, S. (2021). Stress, internet use, substance use and coping among adolescents, young-adults and middle-age adults amid the 'new normal' pandemic era, *Clinical Epidemiology and Global Health, 12 https://doi.org/10.1016/j.cegh.2021.100885*

Snoeren, F., Hoefnagels, C. Measuring perceived social support and perceived stress among primary school children in the netherlands. *Child Indicators Research 7*, 473–486. https://doi.org/10.1007/s12187-013-9200-z

Streiner, D.L., Norman, G.R. & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Şen, S. (2018). Meta-analysis. https://sedatsen.files.wordpress.com/2018/06/meta-analiz.pdf

Şen, S. (2021a). A reliability generalization meta-analysis of Runco ideational behavior scale, *Creativity Research Journal, 34*:2, 178-194, DOI:10.1080/10400419.2021.1960719

Şen, S. (2021b). Current developments in meta-analysis. https://drive.google.com/file/d/1FhCXWjNgmkUSBxi_kZWjJXzAh1v7gfZq/view

Şen, S. & Yıldırım, İ. (2020). *Meta-analysis applications with CMA.* Anı Publishing.

Şeten, C. (2012). *Meta-analysis: an application of the reliability generalization of the multidimensional student life satisfaction scale (mslss).* [Unpublished Master Thesis], Akdeniz University.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics *is* datametrics: The test is not reliable. *Educational and Psychological Measurement, 60,* 174–195.

Toprak, F. (2023). *The mediating role of psychological flexibility in the relationship between obsessive-compulsive symptoms and shame, guilt and perceived stress*. [Unpublished Master Thesis] Yakın Doğu University.

Urbina, S. (2004). *Essentials of psychological testing.* Hoboken, NJ: John Wiley & Sons, Inc.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6-20.

Vacha-Haase, T., Kogan, L.R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those of test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509-522. https://doi.org/10.1177/ 00131640021970682

Vassar, M. & Bradley, G. (2010) A reliability generalization study of coefficient alpha for the life orientation test, *Journal of Personality Assessment, 92*(4), 362-370, DOI: 10.1080/00223891.2010.482016

Vassar, M. & Crosby, J. W. (2008). A reliability generalization study of coefficient alpha for the ucla loneliness scale. *Journal of Personality Assessment. 90*(6), 601-607, DOI: 10.1080/00223890802388624

Wheeler, D. L., Vassar, M., Worley, J. A., & Barnes, L. L. B. (2011). A Reliability Generalization Meta-Analysis of Coefficient Alpha for the Maslach Burnout Inventory. *Educational and Psychological Measurement, 71*(1), 231-244. https://doi.org/10.1177/0013164410391579

Wilkinson, L., & the APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Yerlikaya, E. E. & İnanç, B. (2007). Psychometric properties of the Turkish translation of the perceived stress scale. *IX. National Psychological Counseling and Guidance Congress*, 17-19 October, İzmir.

Yüksel, Ş. (2024). *Factors affecting prenatal perceived stress and prenatal attachment in pregnant women.* [Medical Specialty Thesis]. University of Health Sciences.

Appendix

Altıparmak, E. (2018). *Psikoteknik merkezinde sınava giren bireylerde algılanan stres, stresle başa çıkma tarzları, genel öz yeterlilik ve anksiyete ile sınav başarısı arasındaki ilişki*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Altunkol, F. (2017). *Bilişsel esneklik eğitim programının lise öğrencilerinin bilişsel esneklik ile algılanan stres düzeylerine ve stresle başa çıkma tarzlarına etkisi.* Doktora Tezi, Çukurova Üniversitesi Sosyal Bilimler Enstitüsü, Adana.

Aslan, İ. (2020). *Özel bir temizlik firmasında cilt sorunları sıklığı ve algılanan stres düzeyi ile ilişkisinin incelenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Bal, E. (2020). *Açık kalp ameliyatı olacak hastaların uyku kalitelerinin, kaygı durumlarının ve algıladıkları stres düzeylerinin incelenmesi*. Yüksek Lisans Tezi, İstanbul Okan Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Bardavit, M. (2007). *Kişilik yapılarının- stresi değerlendirme, stresle başa çıkma yaklaşımları, algılanan stres ve iş doyumu üzerinde olan etkisinin karşılaştırmalı olarak incelenmesi*. Yüksek Lisans Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Basım, A. (2016). *Algılanan streste duygusal emek ve kendilik değerlendirmelerinin rolü: avukatlar üzerine bir araştırma*. Yüksek Lisans Tezi, Türk Hava Kurumu Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Başpınar Erten, S. (2020). *Lise son sınıf öğrencilerinin sınav kaygısı ile algılanan stres düzeyleri arasındaki ilişkinin incelenmesi*. Yüksek Lisans Tezi, İnönü Üniversitesi Sağlık Bilimleri Enstitüsü, Malatya.

Bayır, F.N. (2019). *The effect of workplace ostracism on syrian men asylum seekers" perceived stress and subjective well-being*. Yüksek Lisans Tezi, İstanbul Bilgi Üniversitesi Sosyal Bilimler Üniversitesi, İstanbul.

Bülbüloğlu, E. (2018). U*mut, algılanan stres ve baş etme tutumlarının psikolojik belirtilerdeki yordayıcı rolünün incelenmesi.* Yüksek Lisans Tezi, İstanbul Maltepe Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Candan, N. (2019). *Hemşirelerde bilinçli farkındalığa temelli programın algılanan stres ve öz-şefkat üzerine etkisi*. Yüksek Lisans Tezi, İstanbul Okan Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Con Erol, H. (2009). *Kemoterapi alan hastalarda algılanan stresin umutla ilişkisi*. Yüksek Lisans Tezi, Marmara Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Curun, N. (2021). *Üniversitede öğrenim gören sporcuların bilinçli farkındalık düzeyi ve algılanan stres düzeyi arasındaki ilişkinin incelenmesi*. Yüksek Lisans Tezi, Mersin Üniversitesi Eğitim Bilimleri Enstitüsü, Mersin.

Çabuk Aktaş, D. (2021*). Psikolojik danışmanların iş yaşam kalitelerinin, bilişsel duygu düzenleme stratejileri ve algıladıkları stres düzeyleri ile ilişkisinin incelenmesi*. Yüksek Lisans Tezi, Ufuk Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Çabuk, M. (2021). *Psikiyatri servislerinde çalışan hemşirelere göre yöneticilerinin çatışma yönetimi stilleri ve algılanan stres düzeyleri*. Yüksek Lisans Tezi, İstanbul Okan Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Çağlayan, B. (2021). Şizofreni hastalarına uygulanan müzik terapisinin benlik saygısı, motivasyon ve algıladıkları stres düzeyleri üzerine etkisi. Yüksek Lisans Tezi, Çankırı Karatekin Üniversitesi Sağlık Bilimleri Enstitüsü, Çankırı.

Çayırdağ, N. (2012). P*erceived social support, academic self-efficacy and demographic characteristics as predictors of perceived stress among turkish graduate students in the usa.* Doktora Tezi, Orta Doğu Teknik Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Çelik Örücü, M. (2005). *The effects of the stress management training programme on perceived stress, self-efficacy and coping styles of university students*. Doktora Tezi, Orta Doğu Teknik Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Çelik, B. (2016). *Evli bireylerin algıladıkları stres düzeyi ve algıladıkları sosyal destek düzeylerinin evlilik uyumlarına etkisi.* Yüksek Lisans Tezi, Üsküdar Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Çelik, A. T. (2018). *Gündüz ve akşam derslerine devam eden üniversite öğrencilerinin, psikolojik ihtiyaç doyumu, algılanan stres düzeyleri ve akademik erteleme davranışları yönünden karşılaştırılması*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Çobanoğlu, F. N. (2021). *Hekimlerde tükenmişlik düzeyi ve algılanan stres düzeyinin gastrointestinal semptomlar ile ilişkisi.* Uzmanlık Tezi, İstanbul Medeniyet Üniversitesi Göztepe Eğitim ve Araştırma Hastanesi, İstanbul.

Demir Alkin, E. (2019). *Üç ve üzeri gebeliği olan kadınların algıladıkları stres düzeyi ve kendilerini algılama düzeyi ilişkisi.* Yüksek Lisans Tezi, İstanbul Okan Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Demir, T. (2018). *Sağlık çalışanlarında algılanan stres, psikolojik sağlamlık ve bilişsel duygu düzenleme stratejilerinin durumluk ve sürekli kaygı düzeyini yordama gücü.* Yüksek Lisans Tezi, İstanbul Arel Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Demir, M. (2019). *Lise öğrencilerinin problem çözme becerileri, algılanan stres ve yaşam doyumu düzeyleri arasındaki ilişkinin incelenmesi.* Yüksek Lisans Tezi, İstanbul Sabahattin Zaim Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Doğan Bulut, N. (2020). *Evli bireylerin aile yaşam döngüsü basamaklarında algıladıkları stres ve aile stresörleri ile başa çıkma yöntemlerinin incelenmesi.* Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Doğruel, E. (2020). *Tıp fakültesi öğrencilerinde yeme tutumları ile algılanan stres ve stresle başa çıkma arasındaki ilişkinin incelenmesi.* Yüksek Lisans Tezi, Bursa Uludağ Üniversitesi Sağlık Bilimleri Enstitüsü, Bursa.

Duman, A. (2018). *Üst-duyguların, algılanan stres ve psikolojik iyi oluş ile ilişkisinin incelenmesi.* Yüksek Lisans Tezi, İstanbul Maltepe Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Durmuş, E. (2015). *Gebelerin anksiyete, algılanan stres ve depresif belirti durumlarının incelenmesi.* Yüksek Lisans Tezi, İstanbul Medipol Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Efendioğlu, K. (2021). *Otizm spektrum bozukluk tanılı çocuğa sahip ebeveynlerin pandemi sürecindeki başa çıkma stratejileri ve algılanan stres düzeylerinin incelenmesi.* Yüksek Lisans Tezi, Hasan Kalyoncu Üniversitesi Lisansüstü Eğitim Enstitüsü, Gaziantep.

Eren Gülle, T. (2018). *Benlik saygısının, algılanan stres durumu üzerindeki etkisi ve sosyal kaygı ve kaçınmayla ilişkisi.* Yüksek Lisans Tezi, Üsküdar Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Ergüner, B. (2017). *Hekimlerin algıladıkları stres düzeyi, psikolojik dayanıklılıkları ve yaşam doyumları arasındaki ilişkinin incelenmesi*. Yüksek Lisans Tezi, Ufuk Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Güler, B. (2019). *Algılanan stresin yaşam tatminine etkisi*. Yüksek Lisans Tezi, Bolu Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü, Bolu.

Güler, M.Ş. (2019). *Kayak sporcularında kişilik özellikler psikolojik beceri ve algılanan stres arasındaki ilişkinin incelenmesi*. Doktora Tezi, Atatürk Üniversitesi Kış Sporları ve Spor Bilimleri Enstitüsü, Erzurum.

Gülez, A. (2019). *Bipolar bozukluğu olan hastalara verilen stresle başetme eğitiminin stres belirtileri, algılanan stres düzeyi ve stresle baş etme tarzlarına etkisi*. Yüksek Lisans Tezi, Sivas Cumhuriyet Üniversitesi Sağlık Bilimleri Enstitüsü, Sivas.

Gündüz, S. (2019). *Beden eğitimi ve spor öğretmenliği ve sınıf öğretmenliği bölümü öğrencilerinin algılanan stres düzeyleri ile stresle başa çıkma tarzlarının incelenmesi.* Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi Sağlık Bilimleri Enstitüsü, Afyonkarahisar.

Gürbüz, N. B. (2020). Algılanan stres ve uyku kalitesi ilişkisi. Tıpta Uzmanlık Tezi, Sağlık Bilimleri Üniversitesi Ankara Dışkapı Yıldırım Beyazıt Sağlık Uygulama ve Araştırma Merkezi, Ankara.

Hanımoğlu, H. (2014). *Üniversite öğrencilerinde aile fonksiyonları, benliğin ayrımlaşması, algılanan stres, kaygı ve depresyon arasındaki ilişkinin incelenmesi.* Doktora Tezi, Çukurova Üniversitesi Sosyal Bilimler Enstitüsü, Adana.

Havare, G. (2019). *Algılanan stres düzeyinin iş tatmini üzerindeki etkisi: bir kamu hastanesinde çalışan hemşireler üzerinde bir araştırma.* Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Hızıroğlu, Ö. S. (2018). *Spor bilimleri fakültesi öğrencilerinin öznel iyi oluş ve algılanan stres düzeylerinin rekreatif etkinliklere katılım durumları ve farklı değişkenler açısından incelenmesi.* Yüksek Lisans Tezi, Selçuk Üniversitesi Sağlık Bilimleri Enstitüsü, Konya.

Kabahaliloğlu, K. (2019). *Acil ve afetlerde sağlık hizmetleri çalışanlarının algılanan aidiyet, algılanan stres ve problem çözme becerilerinin incelenmesi*. Yüksek Lisans Tezi, İstanbul Gelişim Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Kara, E. (2021). *Öğrenci sporcularda algılanan stres ile psikolojik sağlamlık ilişkisi: başa çıkma stratejileri, bilinçli farkındalık ve algılanan sosyal desteğin aracılığı*. Doktora Tezi, Eskişehir Anadolu Üniversitesi Eğitim Bilimleri Enstitüsü, Eskişehir.

Karaoğlu, K. M. (2016). *Ekonomik güç ve intihar olasılığı arasındaki ilişki: problem çözme becerileri, evlilik uyumu ve algılanan stres açısından bir inceleme*. Yüksek Lisans Tezi, Ankara Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Kaya, M. O. (2019). *Otizmli çocuğu olan annelerde grup rehberliği programının algılanan stres düzeyine etkisi.* Yüksek Lisans Tezi, Bolu Abant İzzet Baysal Üniversitesi Eğitim Bilimleri Enstitüsü, Bolu.

Kayaoğlu, K. (2020). *Alkol ve madde kullanım bozukluğunda bilişsel davranışçı model temelli psikoeğitim destekli müziğin stres, öz yeterlilik ve relasp düzeyine etkisi.* Doktora Tezi, Atatürk Üniversitesi Sağlık Bilimleri Enstitüsü, Erzurum.

Koçyiğit, B. (2018). *Mediating role of perseverative cognition on the relationship between perceived stress and somatic symptoms.* Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Kural, A. I. (2017). *The relationship between unıversity adjustment, attachment style,personality, and perceived stress*. Yüksek Lisans Tezi, Yaşar Üniversitesi Sosyal Bilimler Enstitüsü, İzmir.

Kurt, H. (2021). *Sosyal hizmet kurumlarında çalışan meslek elemanlarının bilinçli farkındalık düzeyleri ile algıladıkları stres düzeyleri ve yaşam doyumları arasındaki ilişkinin incelenmesi*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Küçükant, U. (2019). *X ve y kuşağı çalışanların iş yaşam dengesinin algılanan stres düzeyine etkisi: gıda sektöründe bir işletme araştırması*. Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Leblebicioğlu, M. (2018). *Algılanan stres, bilişsel duygu düzenleme stratejileri ve yeme tutumları arasındaki ilişkinin incelenmesi*. Yüksek Lisans Tezi, İstanbul Maltepe Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Malkoç, D. (2019). *Üniversite öğrencilerinde kendine zarar verme davranışı ile ilişkili faktörlerin incelenmesi: mükemmeliyetçilik, strese yönelik tepkisellik ve baş etme tarzlarının rolü.* Yüksek Lisans Tezi, İstanbul Maltepe Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Malkoç, E. (2021). *Öğretmenlerde rasyonel duygucu öz kararlılık düzeyi, algılanan stres ve iç doyumu arasındaki ilişkinin incelenmesi.* Yüksek Lisans Tezi, Ankara Yıldırım Beyazıt Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Maraqua, A. (2017). *The relationshıp between perceived stress and work engagement: an empirical study.* Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Özbek, H. (2021). *Gebelere uygulanan haptonominin algılanan stres, doğum korkusu ve prenatal bağlanma üzerine etkisi*. Doktora Tezi, Sivas Cumhuriyet Üniversitesi Sağlık Bilimleri Enstitüsü, Sivas.

Özkoçak, E. (2016). *Alzheimer hastalarına bakan kişilerin psikolojik dayanıklılık düzeylerinin algılanan stres, sosyal destek ve demografik özelliklere göre incelenmesi*. Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Öztürk, A. (2019). *Hemşirelerin algıladıkları stres ile premenstural sendrom düzeyleri ilişkisi*. Yüksek Lisans Tezi, İstanbul Okan Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Özyıldırım, E. (2015). *Erzurum il merkezinde çalışan hekimlerde yaşam kalitesi, algılanan stres düzeyi ve etkileyen faktörlerin incelenmesi.* Uzmanlık Tezi, Atatürk Üniversitesi Tıp Fakültesi, Erzurum.

Pehlivan, T. (2019). *Onkoloji-hematoloji hemşirelerine uygulanan kısa ve uzun süreli merhamet yorgunluğu dayanıklılık programının yaşam kalitesi, algılanan stres ve psikolojik dayanıklılık üzerine etkisi.* Doktora Tezi, Koç Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Plakolli, A. (2021). *The roles of perceived organizatıonal support, work-life balance and perceived stress on job performance- a research in kosovo.* Doktora Tezi, Niğde Ömer Halisdemir Üniversitesi Sosyal Bilimler Enstitüsü, Niğde.

Razurel, C., Kaiser, B., Dupuis, M., Antonietti, J. P., Citherlet, C., Epiney, M., & Sellenet, C. (2014). Validation of the Antenatal Perceived Stress Inventory. *Journal of health psychology*, *19*(4), 471–481. https://doi.org/10.1177/1359105312473785

Sheu, S., Lin, H.S. & Hwang, S.L. (2002) Perceived Stress and Physio-Psycho-Social Status of Nursing Students during Their Initial Period of Clinical Practice: The Effect of Coping Behaviors. *International Journal of Nursing Studies, 39*, 165-175. http://dx.doi.org/10.1016/S0020-7489(01)00016-5

Snoeren, F., & Hoefnagels, C. (2014). Measuring Perceived Social Support and Perceived Stress Among Primary School Children in The Netherlands. *Child Indicators Research*, *7*(3), 473-486. https://doi.org/10.1007/s12187-013-9200-z

Süloğlu, D. (2019). *Benlik farklılaşması, algılanan stres düzeyi ve psikolojik sağlamlık arasındaki ilişkinin incelenmesi*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Şahin, G. (2016). *Hemşirelik öğrencilerinde psikolojik sağlamlığın öz yeterlik, sosyal destek ve etkili baş etme ile ilişkisinde algılanan stresin etkisi.* Yüksek Lisans Tezi, İstanbul Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Şahin, B. (2017). *112 acil sağlık hizmetleri çalışanlarında algılanan stres ile saldırganlık arasındaki ilişkilerde sürekli öfke ve algılanan sosyal desteğin aracılık rolü.* Yüksek Lisans Tezi, Ufuk Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Şenocak, S. Ü. (2016). *Hemşirelik öğrencilerinde algılanan stres, algılanan sosyal destek, öznel iyi oluş ve ilişkili faktörler*. Yüksek Lisans Tezi, Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü, Aydın.

Tanlı, A. (2019). *Üniversite öğrencilerinde algılanan stres, umutsuzluk düzeyi ve yeme tutumları arasındaki ilişkilerin incelenmesi*. Yüksek Lisans Tezi, Üsküdar Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Turan, A. (2021). *Covid-19 pandemisi sırasında yüksek risk altında ve düşük risk altında çalışan sağlık çalışanlarının tükenmişlik, kaygı, depresyon, algılanan stres düzeyi, somatik belirtiler, psikolojik dayanıklılık, damgalama ve travmatize olma düzeylerinin karşılaştırılması*. Uzmanlık Tezi, Bolu Abant İzzet Baysal Üniversitesi Tıp Fakültesi, Bolu.

Türkmen, O. O. (2019). *Üniversite öğrencilerinin ölüm kaygısı ile algılanan stres düzeyi arasındaki ilişkide bilinçli farkındalığın aracı rolü*. Yüksek Lisans Tezi, Kırıkkale Üniversitesi Sosyal Bilimler Enstitüsü, Kırıkkale.

Urgancı, Ç. (2018). *İstanbul 112 acil sağlık hizmetleri çalışanlarının mesleki tükenmişlik düzeyleri, algılanan stres düzeyi ve stresle başa çıkma stillerinin evlilik doyumunu yordama gücü.* Yüksek Lisans Tezi, İstanbul Aydın Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Urhan, E. (2020). *The investigation of the link between perceived stress, perfectionism, body image percepiıon and steroid hormones in late adolescent females during the menstrual cycle.* Yüksek Lisans Tezi, Ted Üniversitesi, Lisansüstü Programlar Enstitüsü, Ankara.

Üzbe Atalay, N. (2019). *Lgbt bireylerde kendini toparlama gücü, cesaret, algılanan stres ve sosyal destek*. Doktora Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Yalçın, S. (2013). *İlköğretim okulu öğretmenlerinin mesleki tükenmişlik düzeyleri ile stres, psikolojik dayanıklılık ve akademik iyimserlik arasındaki ilişki*. Yüksek Lisans Tezi, Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Yerlikaya, E. E. (2009). *Üniversite öğrencilerinin mizah tarzları ile algılanan stres, kaygı ve depresyon düzeyleri arasındaki ilişkinin incelenmesi*. Doktora Tezi, Çukurova Üniversitesi Sosyal Bilimleri Enstitüsü, Adana.

Yıldırım, S. (2021). *Palyatif bakım kliniğinde yatan hastaların bakım vericilerinin algılanan stres düzeylerinin ve uyku kalitelerinin belirlenmesi*. Yüksek Lisans Tezi, Erzincan Binali Yıldırım Üniversitesi Sağlık Bilimleri Enstitüsü, Erzincan.

Yıldız, F. (2017). *Mesleki ve teknik anadolu lisesi öğrencilerinin öğrenilmiş güçlülük düzeylerinin algılanan stres düzeyi ile ilişkisinin ve etkileyen diğer faktörlerin incelenmesi*. Yüksek Lisans Tezi, Okan Üniversitesi Sağlık Bilimleri Enstitüsü, İstanbul.

Yılmaztürk, N. H. (2018). *The mediator role of emotion focused coping on the relationship between perceived stress and emotional eating*. Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi Sosyal Bilimleri Enstitüsü, Ankara.

Yılmaztürk, N. (2020). *Algılanan stres, yaşam olayları, psikolojik iyi oluş ve kişilik bozuklukları arasındaki ilişkilerde içsel ve kişiler arası duygu düzenleme güçlüğünün rolünün incelenmesi*. Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Sosyal Bilimleri Enstitüsü, İzmir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 247-259

Discovering Hidden Patterns: Applying Topic Modeling in Qualitative Research

Osman TAT* İzzettin AYDOĞAN**

Abstract

In qualitative studies, researchers must devote a significant amount of time and effort to extracting meaningful themes from large sets of texts and examining the links between themes, which are frequently done manually. The availability of natural language models has enabled the application of a wide range of techniques to automatically detecting hierarchy, linkages, and latent themes in texts. This paper aims to investigate the coherence of the topics acquired from the analysis with the predefined themes, as well as the hierarchy between topics, the similarity, and the proximity-distance between topics by means of the topic model based on BERTopic using unstructured qualitative data. This paper aims to investigate the coherence of the topics acquired from the analysis with the predefined themes, as well as the hierarchy between topics, the similarity, and the proximity-distance between topics by means of the topic model based on BERTopic using unstructured qualitative data. The qualitative data for this study was gathered from 106 students engaged in a university-run pedagogical formation certificate program. In BERTopic procedure, the paraphrase-multilingual-MiniLM-L12-v2 model was used as the sentence transformer model, UMAP was used as the dimension reduction method, and HDBSCAN algorithm as the clustering method. It was found that BERTopic successfully identified six topics corresponding to the six predicted themes in unstructured texts. Moreover, 74% of the texts containing some certain themes could be classified accurately. The algorithm effectively discerned which themes were analogous and which had significant distinctions from others. It was concluded that BERTopic is a procedure which is capable of identifying themes that researchers may not notice, depending on the data density in qualitative data analysis, and has the potential to enable qualitative research to reach more detailed findings.

Keywords: BERTopic, natural language processing, topic modeling

Introduction

Qualitative research, which has an important place in the field of social sciences, is based on a scientific methodology that enables in-depth examination and understanding of the phenomena by analyzing qualitative data obtained from words, pictures, or observations (Chwalisz et al., 1996; Rossman & Rallis, 2017). Researchers often use interviews, focus group discussions, observations, and documents as forms of data collection to draw comprehensive conclusions about the research topic. These techniques are very effective methods for understanding the underlying reasons behind the experiences, perspectives and actions of the participants and the way in which these actions occur (Dinçer & Yavuz, 2023; Wang & Heppner, 2011). In qualitative research, the process of analysis that enables meaningful inferences from the collected data is fundamental. Researchers follow a schematic approach that involves coding, categorizing, and interpreting data to identify patterns, themes, and relationships (Chang & Berk, 2009; Wildemann, 2023). During this iterative analytical process, researchers typically aim to gain a clearer understanding of the research findings by continuously analyzing and comparing their data with the developing conceptual framework (Levitt et al., 2018; Polkinghorne, 1994). The tasks of classifying, categorizing, and extracting relevant patterns and themes from data are complex and need significant attention. These operations are often done manually. However, with the extensive use of natural

* Asst. Prof.,Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, osmantat@yyu.edu.tr, ORCID ID: 0000-0003-2950-9647
** Asst. Prof.,Van Yüzüncü Yıl University, Faculty of Education, Van-Türkiye, izzettinaydogan@yyu.edu.tr, ORCID ID: 0000-0002-5908-1285

To cite this article:
Tat, O., & Aydoğan, İ. (2024). Discovering hidden patterns: Applying topic modeling in qualitative research. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 247-259. https://doi.org/10.21031/epod.1539694
Received: 27.08.2024
Accepted: 17.10.2024

language processing, computers can now perform the difficult task of analyzing qualitative data with a proficiency comparable to that of humans.

The term natural language processing (NLP), which was introduced into our lives through the GPT language model and is rapidly gaining popularity, can be academically described as computers automatically processing both spoken and written human languages. This computational processing of human language enables computers to understand, interpret, and even generate documents or speeches in the target language (Aggarwal & Nair, 2012; Chowdhary, 2020; Pérez-Paredes et al., 2018). NLP includes tasks like part-of-speech tagging, chunking, named entity recognition, language modeling, and semantic role labeling (Tufféry, 2022). Topic modeling, one of the natural language processing methods, is a technique that allows for the detection of latent topics, trends, and themes in large textual datasets known as corpora. This approach is more flexible and effective than conventional techniques, including document clustering (Sudigyo, 2023; Yin & Yuan, 2022), as it allows one to find underlying themes in textual materials. Topic modeling helps researchers expose semantic patterns and structures within textual data, thereby facilitating a better knowledge of the underlying topics found in the data (Boussaadi et al., 2023; Özyurt, 2022; Shin et al., 2023).

Latent Dirichlet allocation (LDA) is the most frequently applied topic modeling technique in machine learning and natural language processing. LDA is a generative probabilistic model that uses unstructured documents as themes and represents each topic through a word distribution (Ekinci & Omurca, 2019; Foster, 2016). While LDA is an effective approach for extracting keywords connected to hidden themes in large collections of papers, the traditional LDA method lacks the capacity to use sentimental meanings during topic extraction (Im et al., 2019). Data sparseness and its incapacity to predict the order of words in documents (Ogunleye et al., 2023) are also known to cause inconsistent outcomes in this method (Weisser et al., 2022). There are also hybrid topic modeling techniques that can effectively overcome the aforementioned inadequacies of LDA. The most important of these is BERTopic, a modern topic modeling technique (Mendonça, 2024; Qiang et al., 2017). BERTopic (Grootendorst, 2022) is an advanced topic modeling technique that generates document embeddings using pre-trained language models (BERT), clusters these embeddings, and finally presents the topics through a class-based TF-IDF procedure. Unlike traditional methods, BERTopic supports multiple topic modeling, hierarchically reduces the number of topics to a fewer number of clusters and automatically determines the appropriate number of topics (Egger & Yu, 2022).

In qualitative studies, researchers must devote significant time and effort to extracting meaningful themes from large sets of texts and examining the links between themes, which are frequently done manually. The availability of natural language models has enabled the application of a wide range of techniques for automatically detecting hierarchies, linkages, and latent themes in texts. It might be argued that although topic modeling with the BERTopic approach has the highest potential among these methods, especially in scientific fields like educational sciences or psychology, where qualitative data is frequently used, its contributions have not yet been fully acknowledged. Therefore, conducting topic modeling based on BERTopic during the analysis of qualitative data is important for revealing possible new approaches that are appropriate to the nature of qualitative research. This paper aims to investigate the coherence of the topics acquired from the analysis with predefined themes, as well as the hierarchy, similarity, and proximity-distance between the topics by means of the topic model based on BERTopic using structured qualitative data. Seeking solutions to four research questions, the study focuses on the advantages of topic modeling in qualitative research. These are:

RQ1: Does the number of extracted topics match the predicted number of topics?
RQ2: To what extent are the main topics extracted consistent with the presumed themes?
RQ3: What is the hierarchy of extracted topics?
RQ4: How is the proximity-distance of the topics?

Literature Review

To the best of our knowledge, few studies examine how topic modeling can be used in qualitative research processes and the various possibilities it presents in terms of its contributions to educational sciences or psychology. Although few studies on topic modeling in the literature on educational sciences exist, it is observed that all of them were carried out using the LDA approach or other probabilistic methods. Çavuşoğlu et al. (2023) carefully examined the English teachers' techno-pedagogical subject knowledge in the Web of Science and Springer databases using the LDA methodology. Foster and Inglis

(2018) also examined the subjects and trends in the complete archive of two academic journals on mathematics education in the UK using the LDA approach.
Bent et al. (2021) examined the feedback written by peers on video recordings of pre-service teachers' teaching experiences with the help of LDA. The outputs of the method enabled them to track the development of pre-service teachers in teaching activities. Similarly, Hujala et al. (2020), using the LDA method, examined students' responses to open-ended feedback questions with the help of topic modeling. Wilson et al. (2024) examined secondary school students' perceptions of automated writing evaluation (AWE) with the help of a topic model based on their responses to open-ended interview questions. In addition to these studies, it is possible to come across many studies conducted with the topic model in the field of educational sciences. LDA algorithm has been used to examine how students access instructional materials without purchasing them (Mosia, 2024), to explore discursive contexts related to social institutions (Soysal & Baltaru, 2021), to analyze the views of faculty staff on the transition from conventional education to online (Casillano, 2022), and to study curriculum texts that define the skills to be taught (Kiener et al., 2023). As can be derived from the literature review, no study discusses how to analyze qualitative data in the field of educational sciences or psychology using BERTopic, an innovative approach that offers significant potential.

Methods

Population and Sample

106 students enrolled in the pedagogical formation certificate program run at Van Yüzüncü Yıl University, Faculty of Education, are the source of qualitative data used in this study. A simple random sampling strategy was used in the data collection process.

Data Collection Tools

The data utilized in this study were collected through online means. In addition to demographic variables, the online form included six open-ended questions that participants answered in writing. These questions were specifically related to six different aspects of the certificate program. The research collected students' views on various aspects of the certificate program, including class size, program planning, course effectiveness, competence of instructors, student-instructor communication, and the quality of measurement and assessment activities.

Data Analysis and Procedure

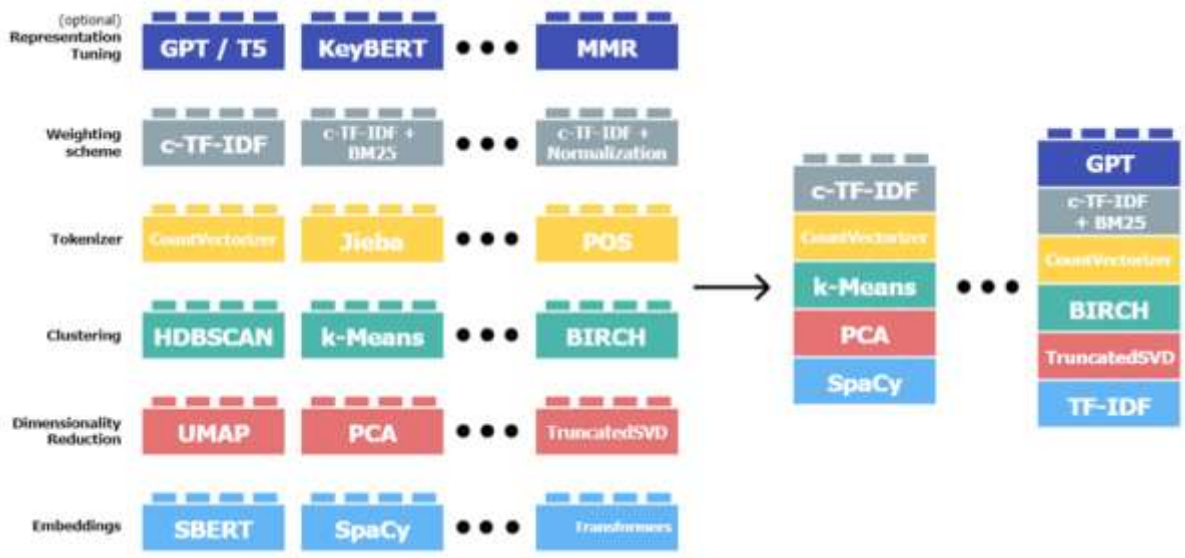
Before the data were analyzed with the topic model, punctuation marks, emojis, and any possible non-textual symbols were removed. Next, stop words that do not contribute to the semantic meaning of the text (e.g., I, me, down, also, etc.) were removed from the text. Afterward, the entire text was converted into lowercase to prevent the same word in uppercase and lowercase from being perceived as different tokens. The texts, consisting of answers to six open-ended questions, were combined into a single column, and the information regarding which text corresponded to which question was hidden from the BERTopic algorithm. When each of the 636 texts collected in a single column is considered as a document, the first step of the BERTopic algorithm is to calculate the embeddings of these documents. In this process, the sentence transformer model ‘paraphrase-multilingual-MiniLM-L12-v2’ (Reimers & Gurevych, 2019) was used. The model, which supports many languages including Turkish, transforms sentences or texts into 384-dimensional dense vector for use in semantic analyses. This methodology was chosen because it was found to produce more consistent outcomes in Turkish texts. UMAP (McInnes et al., 2018) was used to reduce the obtained sentence/text embeddings to a manageable number of dimensions. In the UMAP algorithm, the number of neighbors was set to 10, and the number of components was set to 7. The HDBSCAN (McInnes et al., 2017) algorithm was used to cluster the reduced dimensions by setting the smallest cluster size to 15. As a vectorization method, the count vectorizer was used in a word-based approach and the n-gram parameter was set to (1, 3). The Python libraries sentence transformers, umap, hdbscan, sklearn and BERTopic were used to analyze the data.
Prior to conducting topic modeling on the texts, qualitative analysis was not used to discover the themes. The texts were categorized based on the question to which they were written as a response, and this question served as the theme of the texts. For instance, if a response was provided to the inquiry regarding class size, it was presumed that the content pertained to the theme of class size. The writings

included in the study were not translated from the source language, Turkish, into any other language. Since the analyses were performed on Turkish texts, the visuals and table contents were also presented in Turkish.

BERTopic

BERTopic (Grootendorst, 2022) is a topic modeling method that uses transformer-based language models, namely Bidirectional Encoder Representations from Transformers (BERT), to extract coherent and meaningful topics from textual input. BERTopic differs from standard methods like Latent Dirichlet Allocation (Blei et al., 2003) by using BERT's context-sensitive embeddings to better capture the semantic meaning of words. This leads to more easily understandable subjects (Ding et al., 2023; Hamelberg, 2024). One of the key advantages of BERTopic is its adaptability, which arises from its flexible modeling technique that does not necessitate a predetermined number of topics (Cowan et al., 2022; Yang et al., 2023). BERTopic has additional features such as hierarchical clustering and interactive intertopic distance maps, which facilitate the exploration and analysis of a wide range of topics present in the data (Ramamoorthy, 2024). BERTopic is regarded as an ideal choice for various applications, since research has demonstrated its advantages over other topic modeling methods like as LDA in terms of accuracy and clustering performance (Scarpino et al., 2022; Watanabe & Baturo, 2024). BERTopic can be seen as a series of processes to be taken in order to build topic representations. This approach requires five essential steps to be completed while modeling topics. These steps include embedding generation, dimension reduction, clustering, tokenization, and weighting schemes. The presentation tuning stage is optional. BERTopic is a modular model in which each phase operates independently, and numerous approaches can be applied at each stage. When conducting topic modeling, any of these strategies may be preferred depending on the research goals. For example, HDBSCAN, k-means, and many other clustering algorithms can be used in the clustering stage, whereas GPT, BERT, and many other natural language models can be used in the representation of tuning stage. Thus, each researcher can create a unique BERTopic model tailored to their research (Grootendorst, 2022).

Figure 1.
BERTopic Components



The initial step in BERTopic is to transform input documents into numerical representations in a vector space. At this point, it is assumed that documents with the same topic exhibit semantic similarity. BERTopic employs the Sentence-BERT (SBERT) structure developed by Reimers and Gurevych (2019). This framework efficiently transforms words and paragraphs into vectors using pre-trained language models (Bianchi et al., 2020). Another key component of BERTopic is the process of reducing

the dimensionality of input embeddings. This is because embeddings with high dimensionality create challenges for clustering. One possible option is to reduce the dimensionality of the embeddings to a manageable dimensional space that can be used by clustering methods. As a dimension reduction technique, UMAP is employed as the default method in BERTopic because it has the ability to represent both the local and global high-dimensional spaces in lower dimensions (McInnes et al., 2018). Once the dimensionality of the input embeddings has been reduced, it is necessary to cluster them into groups of comparable embeddings in order to extract topics. The process of clustering is essential, as the effectiveness of clustering technique directly impacts the accuracy of topic representations. HDBSCAN is the typical choice for clustering in BERTopic. Thanks to its ability to effectively capture structures with varying densities (Wang et al., 2021; Zhang et al., 2018), this technique ensures that unrelated documents are not assigned to any cluster, thus enhancing the quality of topic representations.

The interpretability of topic representations plays an important role in topic modeling. Topic representations are generated based on the distribution of texts in each cluster and assigned to a single topic. This process depends on the identification of distinguishing factors between topics based on the word distributions in the clusters. In BERTopic, this process is performed by cluster-based Term Frequency-Inverse Document Frequency (c-TF-IDF), an adjusted version of the TF-IDF (Term Frequency and Inverse Document Frequency) metric, which measures the importance of a word in a document (Egger & Yu, 2022). In this method, instead of being represented as a set of articles, each cluster is transformed into a single document. Then, the frequency of word x in cluster c is calculated, where c corresponds to the previously created cluster. This results in a class-based term frequency representation. Finally, this representation is L1-normalised to adjust for differences in topic sizes (Grootendorst, 2022). For a term x within class c, c-TF-IDF.

$$W_{x,c} = \| tf_{x,c} \| \times \log(1 + \frac{A}{f_x}) \quad (1)$$

$tf_{x,c}$: frequency of word $\times$ in class c

f_x: frequency of word $\times$ across all classes

A: average number of words per class

Results

As a result of the unsupervised topic modeling, it was concluded that the unstructured text contains eight latent topics. The BERTopic algorithm assigns a value of -1 to any topic considered an outlier, which is not analyzed or interpreted as part of the analysis. In this scenario, it is more appropriate to analyze and comment on the details of the seven topics derived from the model. The results are displayed in Table 1 and Figure 2.

Figure 2.

Topic Word Score

Table 1

Topics and Representations

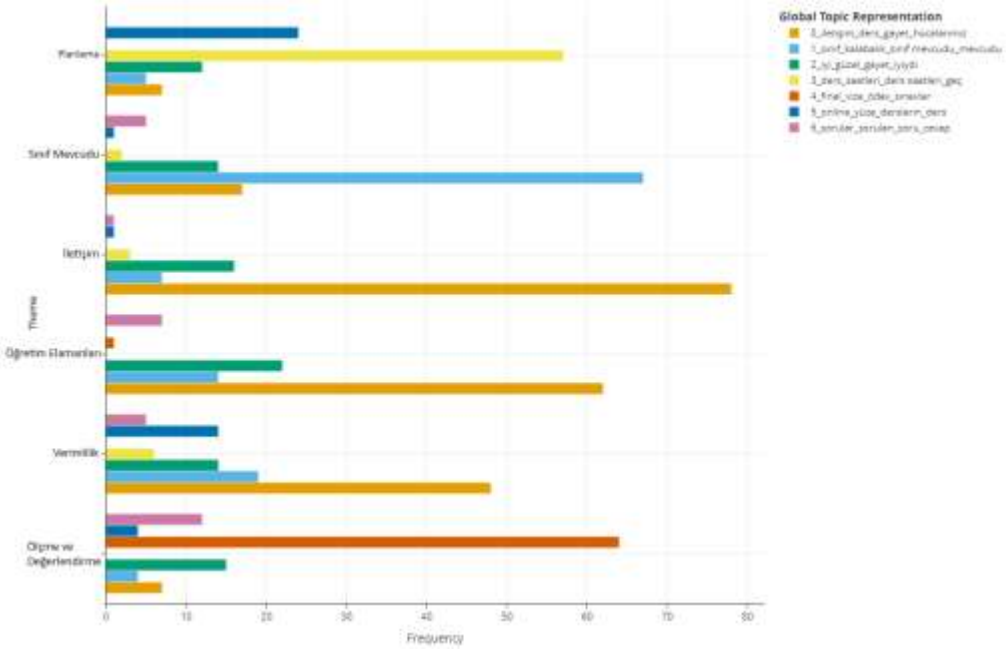
Topic	Count (n)	Name	Token/n-gram	Representative_Docs
-1	1	-1_değildir_yeterli_ doğru karar_karar	['değildir', 'yeterli', 'doğru karar', 'karar', 'dönemde', 'doğru', 'öğretmenlik', 'olağan', 'sorun yaşadım', 'varken formasyon']	"Çünkü 4 yıl eğitim görmek varken formasyon ile iki dönemde alıyoruz, bu yeterli derecede maalesef değildir. Fakat hocalarımızdan aldığımız bilgiler umarım bizler için yeterli olur çünkü buna öğretmenlik mesleğine başlamadan net bir şey söylemiyorum. Fakat yeterli değildir."
0	219	0_iletişim_ders_gayet_ hocalarımız	['iletişim', 'ders', 'gayet', 'hocalarımız', 'dersler', 'öğretim', 'iyi', 'verimli', 'akademik', 'hocalarımızın']	"Ders dışında ders ile alakalı olarak ihtiyaç duyulduğunda sınıf temsilcileri ile iletişim kurulması biraz garip olsa da gerekli açıklamalar kulaktan kulağa yapılmaktadır. Ders sırasında herhangi bir iletişim sorunu olmadan uygun açıklamalar yapılıyor."
1	116	1_sınıf_kalabalık_ sınıf mevcudu_	['sınıf', 'kalabalık', 'sınıf mevcudu', 'mevcudu', 'fazla', 'olumsuz', 'öğrenme', 'öğrenci', 'değildi', 'sınıf mevcudunun']	"Sınıf mevcudu ortalama 45 kişilik. Öğrenme sürecini olumsuz etkilediğini düşünmüyorum çünkü genel olarak herkesin yaşı itibari ile farkındalığından sınıf düzeni bozulmuyor çok fazla."
2	93	2_iyi_güzel_gayet_ iyiydi	['iyi', 'güzel', 'gayet', 'iyiydi', 'verimli', 'yok', 'yeterli', 'uygun', 'gerektiği', 'gayet iyi']	"İyi", "iyi", "iyi"
3	68	3_ders_saatleri_ders saatleri_geç	['ders', 'saatleri', 'ders saatleri', 'geç', 'akşam', 'gün', 'derslerin', 'saatler', 'sayısı', 'saatte']	"Örgün eğitim gören öğrenciler olduğundan dolayı formasyon öğrencilerinin akşam saatlerinde ders görüyor olmaları normal fakat ders saatleri olarak bakıldığında yeterli değildi. Öğretmenlik uygulaması ise benim açımdan verimli ve bilgi edindiğim bir program oldu."
4	65	4_final_vize_ödev_ sınavlar	['final', 'vize', 'ödev', 'sınavlar', 'sınav', 'ölçme', 'değerlendirme', 'vize final', 'ölçme değerlendirme', 'zor']	"Hocamız bu konu da vize ve finalde soru dağılımlarını çok iyi şekilde ayarlayarak şans başarısını düşürerek güvenilir ve tutarlı bir ders ve sınav yaşattı. Son olarak bu dersin bir dönem de aşılmasının doğru olmadığını da beyan ederim. Yani 4 yıl da alınmalıdır"
5	44	5_online_yüze_ derslerin_ders	['online', 'yüze', 'derslerin', 'ders', 'derslerin online', 'öğretmenlik', 'verimli', 'dersler', 'ders saatleri', 'saatleri']	"Ders saatleri uygun. Online dersleri pek etkili değil yüz yüze olması daha mantıklı. Online olan derslerin sınavları da online olması lazım.", "Bazı derslerin online olmasından çok memnunum. Hepsi online olsa çok daha iyi olur."
6	30	6_sorular_sorulan_ soru_cevap	['sorular', 'sorulan', 'soru', 'cevap', 'güzel', 'bence', 'olabilirdi', 'sorulan sorular', 'gayet', 'işlediğimiz']	"Hocalarımız sorulan sorulara eksiksiz cevap veriyor kafamızda soru işareti kalmıyor', "Hocalarımızın bilgi düzeyleri gayet makuldü konulara hakimiyet sorulan sorular karşısında verdikleri cevap tatmin edici."

The first noteworthy finding is that Topic 2 was categorized as a separate topic because the algorithm could not associate the answers given to all questions with one or two words such as 'good', 'very good', or 'sufficient' with other topics. Since it is not possible to understand and interpret the context from this topic, it does not provide a significant finding for the research. When this topic is neglected, it can be asserted that the other topics precisely coincide with the pre-determined themes. Specifically, Topic 0 relates to the quality of communication between students and instructors, Topic 1 to the size of the class, Topics 3 and 5 to the planning and efficiency of courses, Topic 4 to assessment and evaluation activities,

and Topic 6 to the competence of instructors. Figure 3 presents the specific relationship between the texts related to the topics and the presumed themes.

Figure 3

Topic per Theme



The algorithm classified 74% (n=78) of the texts expected to belong to the communication theme as Topic 0. Similarly, 63% (n=67) of the texts belonging to the theme of class size were classified as Topic 1. 54% (n=57) of the texts known to belong to the theme of lesson planning were classified as Topic 3. 60% (n=64) of the texts known to belong to the theme of measurement and evaluation were classified as Topic 4. 13% (n=14) of the views on the efficiency of the courses were associated with Topic 5. Finally, 7% (n=7) of the texts belonging to the theme of the competence of the lecturers were classified as Topic 6. To examine the extent to which the texts are related to the topics, the BERTopic algorithm provides the classification probability of each text, which is a very useful feature. The average classification probability for each topic indicates how reliably the algorithm classifies each text. Accordingly, the average probability is 0.74 for Topic 0, 0.72 for Topic 1, 0.49 for Topic 2, 0.81 for Topic 3, 0.87 for Topic 4, 0.85 for Topic 5, and 0.64 for Topic 6. In this case, it is possible to say that the algorithm is not reliable enough when the texts are categorized into Topic 2, which consists of short answers such as ‘good’, ‘nice’, and so on.

Figure 4

Similarity Matrix

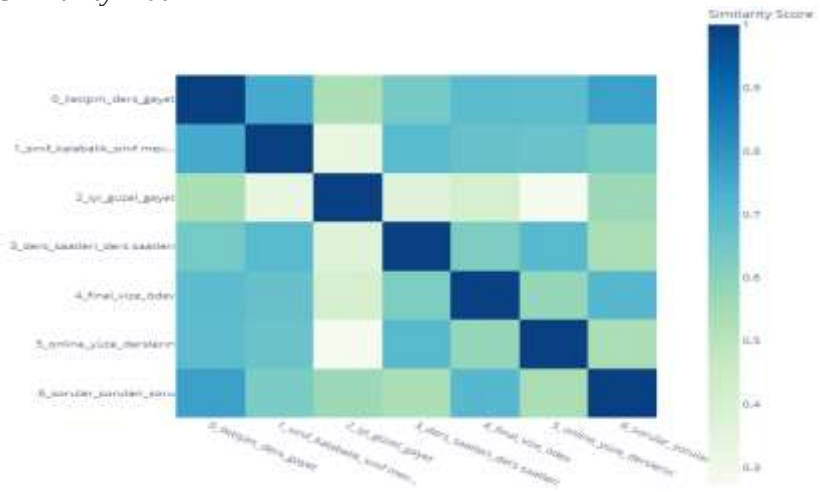
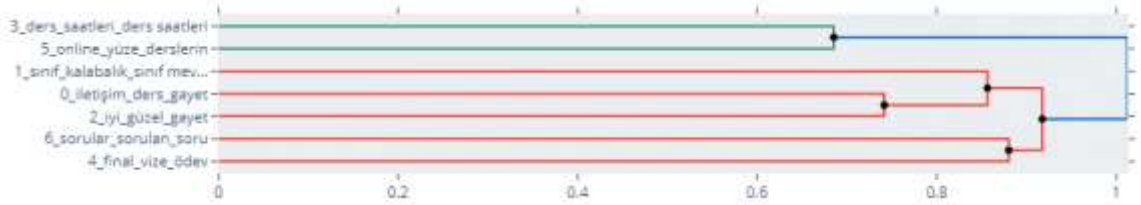


Figure 5

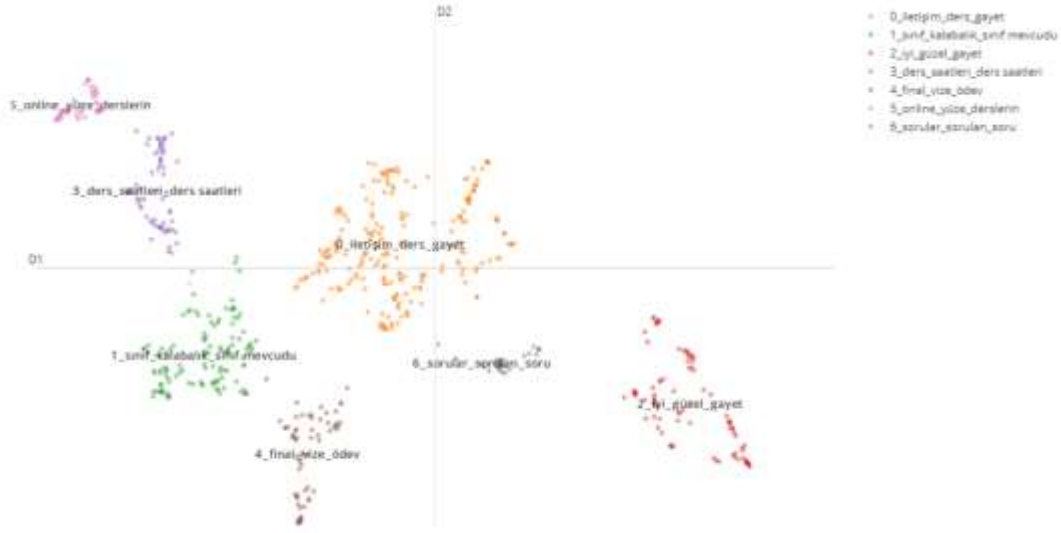
Hierarchical Clusters



One of the most important features of BERTopic for researchers is that the relationship (Figure 4) and hierarchy (Figure 5) between topics can be obtained after topic modeling. When the similarity matrix in Figure 3 and the hierarchy dendrogram in Figure 4 are analyzed, it can be said that there is a strong relationship between Topic 0 (communication theme) and Topic 1 (class size theme). The main reason for this is the large number of opinions in many texts stating that crowded classes make communication difficult, or that classes are not crowded and thus do not negatively affect communication. Similarly, it is possible to mention a strong relationship between Topic 0 and Topic 6. In most of the texts, students stated that they received satisfactory answers to the questions they asked to the lecturers both in and outside of class and that communication was at a sufficient level in this regard. Another notable relationship is observed between Topic 6 and Topic 4. The likely reason for this relationship is that some of the texts classified in Topic 6 are related to the questions asked to the instructors in the classroom, while some of them are related to the questions asked in the exams. Again, it is seen that Topic 3 and Topic 5 have a high level of similarity. This indicates that the themes of the time interval of the courses and the way the courses are organized (online or face-to-face) have many common aspects.

Figure 6

Document Distributions Based on Reduced Embeddings



Another important feature of topic modeling based on the BERTopic algorithm is that a scatter plot can be created from the reduced embeddings obtained from sentence transformers, which are reduced as a result of UMAP and clustering analysis. This graph visually illustrates how the topics are distributed on the two-dimensional surface and how the level of proximity-distance between the topics is. In Figure 6, it is firstly observed that the texts belonging to Topic 2 are quite distant from the other topics and are an isolated cluster with a weak relationship to the others. Again, it is seen that Topic 0, which contains the theme of the quality of communication, is in a more central position and is located especially close to Topic 1 and Topic 6. Similarly, it can be said that Topic 3 and Topic 5 are positioned very closely to each other.

Discussion

This work presents the advantages of utilizing the BERTopic (Grootendorst, 2022) method for extracting topics from unstructured texts. The algorithm integrates many components of natural language models, sophisticated dimensionality reduction, and clustering analysis to provide a more efficient approach to topic modeling (Abuzayed & Al-Khalifa, 2021; Kukushkin, 2022; Maryanto, 2024). As a result of the analysis, the algorithm detected that the texts contained seven hidden topics, except for the outlier topic (Topic -1). One of these seven topics (Topic 2) was found to contain very short answers such as 'good', 'good', 'good', 'enough', and its relationship with the other topics could not be fully determined due to the insufficient token diversity. Thus, it can be said that the topic model successfully identified six topics corresponding to the six predicted themes in unstructured texts. Topic 0 was found to represent the majority of the texts belonging to the theme of the quality of communication between teaching staff and students. Similar situations are also valid for other topics. For example, Topic 1 overlaps with the theme of class size, Topic 3 and Topic 5 with the themes of lesson planning and efficiency of lessons, Topic 4 with the theme of assessment and evaluation activities, and Topic 6 with the theme of competence of lecturers. The accuracy rate of the texts in the topics varies between 78% (Topic 0) and 7% (Topic 6). The lower-than-expected accuracy rates in some topics may be due to many texts containing opinions on more than one theme and a significant portion of the texts consisting of just one or two words. Texts expressing thoughts on multiple themes may have been assigned to the wrong topic, or texts with too few tokens may have been classified under another topic. Still, it can be argued that when the thoughts are detailed and the interview questions are effectively crafted, the BERTopic method can be highly successful.

Another advantage of topic modeling with BERTopic, as revealed in this study, is the ability to identify similarity levels between topics and the topic hierarchies obtained based on these similarity levels (Cheddak, 2024; Kousis, 2023). In this study, the level of similarity between topics was analyzed using a heat map based on the correlation between topics and a scatter plot of the texts generated from reduced embeddings. In light of these analyses, it was possible to determine which texts are more centrally located, i.e. which texts are likely to contain other topics, based on the relationship between topics. Additionally, it was also possible to determine which topics were clearly differentiated from other topics. To determine the hierarchy between texts, the results obtained from the cluster analysis can be analyzed with the help of dendrograms. It can be said that the intertextual hierarchy has many advantages for researchers. For example, identifying some topics as a whole of more than one sub-topic can enhance the researchers' understanding of the relationships between themes and thematization processes while conducting qualitative analyses. In conclusion, BERTopic is an advanced modern technique that combines many useful aspects of natural language processing, dimension reduction and cluster analysis techniques. It is thought that this algorithm will contribute to the automatic mining of qualitative data, which is frequently used in social sciences such as education and psychology, primarily by providing quantitative indicators based on qualitative data, and performing similarity and hierarchical structure analyses based on these quantitative values. In addition, BERTopic can identify themes that researchers may not have noticed, depending on the density of the data in qualitative data analysis, and enable qualitative research to reach more detailed findings.

Declarations

Conflict of Interest: The authors have no conflicts of interest to declare.

Ethical Approval: The study was ethically approved by the by Van Yüzüncü Yıl University Social and Human Sciences Ethics Committee dated 07/08/2024 and numbered 2024/16-17**.**

A part of this study was presented as an oral presentation at Vizyon Van 2024 Congress, Van, Turkey, 2024.

References

Abuzayed, A., & Al-Khalifa, H. S. (2021). Bert for Arabic topic modeling: An experimental study on BERTopic technique. *Procedia Computer Science*, 189, 191-194. https://doi.org/10.1016/j.procs.2021.05.096

Aggarwal, E., & Nair, S. (2012). NLP token matching on database using binary search. *International Journal of Computers & Technology*, 3(1), 140-143. https://doi.org/10.24297/ijct.v3i1c.2766

Bent, M., Velazquez-Godinez, E., & Jong, F. (2021). Becoming an expert teacher: Assessing expertise growth in peer feedback video recordings by lexical analysis. *Education Sciences*, 11(11), 665. https://doi.org/10.3390/educsci11110665

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics*: *Main volume*,1676–1683. doi:10.18653/v1/2021.eacl-main.143

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(1), 993–1022.

Boussaadi, S., Aliane, H., & Abdeldjalil, O. (2023). Using an explicit query and a topic model for scientific article recommendation. *Education and Information Technologies*, 28(12), 15657-15670. https://doi.org/10.1007/s10639-023-11817-2

Casillano, N. F. B. (2022). Discovering sentiments and latent themes in the views of faculty members towards the shift from conventional to online teaching using VADER and latent dirichlet allocation. *International Journal of Information and Education Technology*, 12(4), 290-298. https://doi.org/10.18178/ijiet.2022.12.4.1617

Çavuşoğlu, D., Kıncal, R. Y., & Kartal, O. Y. (2023). Systematic review of research conducted on the techno-pedagogical content knowledge of English teachers. *Journal of Family Counseling and Education*, 8(2), 170-192. https://doi.org/10.32568/jfce.1269034

Chang, D. F., & Berk, A. (2009). Making cross-racial therapy work: A phenomenological study of clients' experiences of cross-racial therapy. *Journal of Counseling Psychology*, 56(4), 521-536. https://doi.org/10.1037/a0016905

Cheddak, A. (2024). BERTopic for enhanced idea management and topic generation in brainstorming sessions. *Information*, 15(6), 365. https://doi.org/10.3390/info15060365

Chowdhary, K. R. (2020). Natural language processing. Fundamentals of Artificial Intelligence, 603-649. https://doi.org/10.1007/978-81-322-3972-7_19

Chwalisz, K., Wiersma, N., & Stark-Wroblewski, K. (1996). A quasi-qualitative investigation of strategies used in qualitative categorization. *Journal of Counseling Psychology*, 43(4), 502-509. https://doi.org/10.1037/0022-0167.43.4.502

Cowan, T., Rodriguez, Z., Granrud, O., Masucci, M., Docherty, N., & Cohen, A. (2022). Talking about health: A topic analysis of narratives from individuals with schizophrenia and other serious mental illnesses. *Behavioral Sciences*, 12(8), 286. https://doi.org/10.3390/bs12080286

Dinçer, P., & Yavuz, H. (2023). Behind the screen: a case study on the perspectives of freshman EFL students and their instructors. *Education and Information Technologies*, 28(9), 11881-11920. https://doi.org/10.1007/s10639-023-11661-4

Ding, Q., Ding, D., Wang, Y., Guan, C., & Ding, B. (2023). Unraveling the landscape of large language models: A systematic review and future perspectives. *Journal of Electronic Business & Digital Economics*, 3, 3-19. https://doi.org/10.1108/jebde-08-2023-0015

Egger, R., & Yu, J. (2022). A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify twitter posts. *Frontiers in Sociology*, 7. https://doi.org/10.3389/fsoc.2022.886498

Ekinci, E., & Omurca, S. (2019). Concept-LDA: Incorporating Babelfy into LDA for aspect extraction. *Journal of Information Science*, 46(3), 406-418. https://doi.org/10.1177/0165551519845854

Foster, A. (2016). An extension of standard latent dirichlet allocation to multiple corpora. *SIAM Undergraduate Research Online*, 9. https://doi.org/10.1137/15s014599

Foster, C., & Inglis, M. (2018). Mathematics teacher professional journals: What topics appear and how has this changed over time?. International Journal of Science and Mathematics Education, 17(8), 1627-1648. https://doi.org/10.1007/s10763-018-9937-4

Grootendorst, M. (2022). BERTOPIC: Neural topic modeling with a class-based TF-IDF procedure. https://doi.org/10.48550/arxiv.2203.05794

Hamelberg, K., de Ruyter, K., van Dolen, W., & Konuş, U. (2024). Finding the right voice: How CEO communication on the Russia–Ukraine war drives public engagement and digital activism. *Journal of Public Policy & Marketing*. https://doi.org/10.1177/07439156241230910

Hujala, M., Knutas, A., Hynninen, T., & Arminen, H. (2020). Improving the quality of teaching by utilizing written student feedback: A streamlined process. *Computers & Education*, 157, 103965. https://doi.org/10.1016/j.compedu.2020.103965

Im, Y., Park, J., Kim, M., & Park, K. (2019). Comparative study on perceived trust of topic modeling based on affective level of educational text. *Applied Sciences*, 9(21), 4565. https://doi.org/10.3390/app9214565

Kiener, F., Gnehm, A., & Backes-Gellner, U. (2023). Noncognitive skills in training curricula and nonlinear wage returns. *International Journal of Manpower*, 44(4), 772-788. https://doi.org/10.1108/ijm-03-2022-0119

Kousis, A. (2023). Investigating the key aspects of a smart city through topic modeling and thematic analysis. *Future Internet*, 16(1), 3. https://doi.org/10.3390/fi16010003

Kukushkin K., Ryabov Y., & Borovkov A. (2022). Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling. *Data*, 7(12):173. https://doi.org/10.3390/data7120173

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. *American Psychologist*, 73(1), 26-46. https://doi.org/10.1037/amp0000151

Maryanto, M. (2024). Hybrid model for extractive single document summarization: Utilizing bertopic and bert model. *IAES International Journal of Artificial Intelligence* (Ij-Ai), 13(2), 1723. https://doi.org/10.11591/ijai.v13.i2.pp1723-1731

McInnes, L., Healy, J. J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29), 861.

Mendonça, M. (2024). Topic extraction: BERTopic's insight into the 117th congress's twitterverse. *Informatics*, 11(1), 8. https://doi.org/10.3390/informatics11010008

Mosia, M. (2024). Data-driven insights into non-purchasing behaviours through latent dirichlet allocation: Analysing study material acquisition among university students. *Journal of Culture and Values in Education*, 7(1), 72-82. https://doi.org/10.46303/jcve.2024.5

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of topic modelling approaches in the banking context. *Applied Sciences*, 13(2), 797. https://doi.org/10.3390/app13020797

Özyurt, Ö. (2022). Empirical research of emerging trends and patterns across the flipped classroom studies using topic modeling. *Education and Information Technologies*, 28(4), 4335-4362. https://doi.org/10.1007/s10639-022-11396-8

Pérez-Paredes, P., Guillamón, C. O., & Jiménez, P. A. (2018). Language teachers' perceptions on the use of oer language processing technologies in mall. *Computer Assisted Language Learning*, 31(5-6), 522-545. https://doi.org/10.1080/09588221.2017.1418754

Polkinghorne, D. E. (1994). Reaction to special section on qualitative research in counseling process and outcome.. Journal of Counseling Psychology, 41(4), 510-512. https://doi.org/10.1037//0022-0167.41.4.510

Qiang, J., Chen, P., Wang, T., & Wu, X. (2017). Topic modeling over short texts by incorporating word embeddings. *Advances in Knowledge Discovery and Data Mining*, 363-374. https://doi.org/10.1007/978-3-319-57529-2_29

Ramamoorthy, T., Kulothungan, V., & Mappillairaju, B. (2024). Topic modeling and social network analysis approach to explore diabetes discourse on twitter in India. *Frontiers in Artificial Intelligence*, 7. https://doi.org/10.3389/frai.2024.1329185

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Retrieved from http://arxiv.org/abs/1908.10084

Reimers, N., & Gurevych, I. (2019). Sentencebert: Sentence embeddings using siamese BERTnetworks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Procesessing Association for Computational Linguistics*.

Rossman, G., & Rallis, S. F. (2017). *An introduction to qualitative research: Learning in the field*. SAGE Publications. https://doi.org/10.4135/9781071802694

Scarpino, I., Zucco, C., Vallelunga, R., Luzza, F., & Cannataro, M. (2022). Investigating topic modeling techniques to extract meaningful insights in italian long covid narration. Biotech, 11(3), 41. https://doi.org/10.3390/biotech11030041

Shin, M., Ok, M. W., Choo, S., Hossain, G., Bryant, D. P., & Kang, E. (2023). A content analysis of research on technology use for teaching mathematics to students with disabilities: Word networks and topic modeling. *International Journal of STEM Education*, 10(1). https://doi.org/10.1186/s40594-023-00414-x

Soysal, Y., & Baltaru, R. (2021). University as the producer of knowledge, and economic and societal value: The 20th and twenty-first century transformations of the UK higher education system. *European Journal of Higher Education*, 11(3), 312-328. https://doi.org/10.1080/21568235.2021.1944250

Sudigyo, D., Hidayat, A. A., Nirwantono, R., Rahutomo, R., Trinugroho, J. P., & Pardamean, B. (2023). Literature study of stunting supplementation in Indonesian utilizing text mining approach. *Procedia Computer Science*, 216, 722-729. https://doi.org/10.1016/j.procs.2022.12.189

Sutton, J., & Austin, Z. (2015). Qualitative research: Data collection, analysis, and management. *The Canadian Journal of Hospital Pharmacy*, 68(3). https://doi.org/10.4212/cjhp.v68i3.1456

Tufféry, S. (2022). *Deep learning: From big data to artificial intelligence with r*. John Wiley & Sons Ltd. https://doi.org/10.1002/9781119845041.ch9

Wang, L., Chen, P., Chen, L., & Mou, J. (2021). Ship AIS trajectory clustering: An HDBSCAN-based approach. *Journal of Marine Science and Engineering*, 9(6), 566. https://doi.org/10.3390/jmse9060566

Wang, Y., & Heppner, P. P. (2011). A qualitative study of childhood sexual abuse survivors in Taiwan: Toward a transactional and ecological model of coping. *Journal of Counseling Psychology*, 58(3), 393-409. https://doi.org/10.1037/a0023522

Watanabe, G., Conching, A., Nishioka, S. T., Steed, T., Matsunaga, M., Lozanoff, S.,…& Noh, T. (2023). Themes in neuronavigation research: A machine learning topic analysis. *World Neurosurgery: X*, 18, 100182. https://doi.org/10.1016/j.wnsx.2023.100182

Watanabe, K., & Baturo, A. (2024). Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences. Social Science Computer Review, 42(1), 224-248. https://doi.org/10.1177/08944393231178605

Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T., … & Säfken, B. (2022). Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using twitter data. *Computational Statistics*, 38(2), 647-674. https://doi.org/10.1007/s00180-022-01246-z

Wildemann, S. (2023). Bridging qualitative data silos: The potential of reusing codings through machine learning based cross-study code linking. S*ocial Science Computer Review*, 42(3), 760-776. https://doi.org/10.1177/08944393231215459

Wilson, J., Zhang, S., Palermo, C., Cordero, T. C., Zhang, F., Myers, M. C., … & Coles, J. (2024). A latent dirichlet allocation approach to understanding students' perceptions of automated writing evaluation. *Computers and Education Open*, 6, 100194. https://doi.org/10.1016/j.caeo.2024.100194

Yang, L., Shi, J., Zhao, C., & Zhang, C. (2023). Generalizing factors of covid-19 vaccine attitudes in different regions: A summary generation and topic modeling approach. *Digital Health*, 9. https://doi.org/10.1177/20552076231188852

Yin, B., & Yuan, C. (2022). Detecting latent topics and trends in blended learning using LDA topic modeling. *Education and Information Technologies*, 27(9), 12689-12712. https://doi.org/10.1007/s10639-022-11118-0

Zhang, D., Lee, K., & Lee, I. (2018). Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. *Expert Systems with Applications*, 92, 1-11. https://doi.org/10.1016/j.eswa.2017.09.040