Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *İlkbahar* 2024 *Spring* 2024 *Cilt: 15-Sayı: 1* *Volume: 15-Issue: 1*

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.)
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.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2024, Sayı: 15-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2024, Issue: 15-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Editorial; 2024; 15(1); 1-4

EDITORIAL

Adapting the Person Fit Analysis: Ideas on Detecting Person Misfit in Computerized Adaptive Testing

Beyza Aksu DÜNYA*

Highlights

- Most test accountability stakes occur at the individual level (Walker & Engelhard, 2016) so person fit analysis is an important part of documenting validity evidence.
- Much of the available research on person fit in Computerized Adaptive Testing (CAT) utilized traditional person fit statistics for detecting person misfit.
- Among the studied approaches, cumulative sum (CUSUM) procedures have been found powerful in CAT but when the parameters of the underlying statistical model are known before and after the change in response string (which doesn't hold in most CAT applications).
- A comprehensive approach with multiple indicators of person fit may be needed.

In this editorial chapter, I aim to summarize findings on person fit analysis in computerized adaptive testing (CAT) from prior research and discuss potential avenues for further research. In item response theory (IRT) applications, person fit quantifies fit of a response pattern to the model (Bradlow & Weiss, 2001, p. 86). Person misfit refers to unexpected response patterns by individuals. There are many potential reasons of misfit including special knowledge (Sinharay, 2016), cheating, guessing (Meijer, 1996), fatigue (Swearingen, 1998), warming up (Meijer, 1996), or faking (Ferrando & Anguiano-Carrasco, 2012). Evaluation of misfit is a significant step for addressing discrepancies within the measurement model. When IRT models are used, evidence of model fit which involves person fit analysis results should be reported (Standard 4.10; AERA, APA & NCME, 2014) as validity evidence to enhance score interpretations. Once misfitting items are identified, corrective steps such as item revision or removal can be implemented. For examinees who exhibit misfit, additional exploration can be undertaken to pinpoint behaviors that might necessitate adjustments to the test program or corrective interventions for particular examinees.

Although IRT estimates are robust to model-data misfit and many control mechanisms, involving both statistical (i.e., standardized log-likelihood index) and graphical approaches (i.e., person response plots), are available to detect person misfit, respondents in real test administrations may respond to items in unique and unstudied way (Walker & Engelhard, 2016). In addition, available misfit measures are specifically designed for fixed-item tests and have lower power when used with adaptive testing (van Krimpen-Stoop & Meijer, 1999, Meijer & van Krimpen-Stoop, 2010, Robin, 2002). This comes from two advantageous features of CAT that is item selection mechanisms which result in shorter tests and

* Assoc. Prof. Dr., Bartın University, Faculty of Education, Bartın-Turkiye, baksu@bartin.edu.tr, ORCID ID: 0000-0003-4994-1429

To cite this article:
Dünya, B.A. (2024). Adapting the person fit analysis: ideas on detecting person misfit in computerized adaptive testing, *15*(1), 1-4. https://doi.org/10.21031/epod.1461703

modest spread of item difficulties for an examinee (Meijer & van Krimpen-Stoop, & E.M.L.A. 2009, p. 32). In CAT, an item selection mechanism based on maximum information is utilized as part of the testing algorithm. This algorithm chooses items from an item pool that best match to the examinee's ability level. It aims to minimize the administration of items that are significantly too easy or too difficult for that examinee. Consequently, every examinee is presented with a unique test comprising items that are targeting for the examinee's ability level. Paradoxically, adaptive nature of CAT reduces the traditional sources of person misfit, while it poses a challenge for the detection of person misfit. In CAT, likelihood of inappropriate item selection that is too hard or too easy for a particular respondent is minimized. However, a person's responses should still be checked for fit to the IRT model chosen to calibrate parameters. Since different sets of items are drawn from an item pool with item parameters considered to be known, person fit checks in CAT, which may be absent in the item pool development stage, should provide additional quality check for data-model fit (Walker and Engelhard, 2016).

To address this concern attached to CAT applications, researchers have developed adaptive test specific person fit statistics and tested their misfit detection power (Hendravan, Glas & Meijer, 2005; McLeod & Lewis, 1999; van Krimpen-Stoop, 2000). A handful person fit indices that performed well in CAT depend on the CUSUM approach (i.e., LARD by Bradlow and Weiss, 2001; iterative upper and lower CUSUM by van Krimpen-Stoop & Meijer; 2000, 2001, 2002). This approach was found particularly successful at identifying abrupt shifts in response patterns, attributed to issues like decreased attention, speededness, or item preknowledge. CUSUM-based statistical process control mechanisms are found the most useful especially when the parameters of the underlying model before and after the change are known (Montgomery, 2013), which is not the case for CAT. Researchers addressed this shortcoming of CUSUM-based fit statistics for detecting person fit by proposing change-point based fit statistics (Tests for change point- TFCP;). Similar to the CUSUM approach, the logic of tests for change point (TFCP) is to find the point where the model parameters underlying a sequence of responses have changed in some fashion. This approach was tested for its usefulness for CAT since item parameters within an item pool are assumed to be known, whereas person parameters are not (Sinharay, 2016). Although TFCP-based fit statistics were found powerful in detecting unexpectedly abrupt change in response string, potential reasons of person misfit is not limited to this in CAT. An abrupt change in response strings can occur due to various reasons, such as initial warming up, speededness/fatigue or loss of attention through the end, or specialized content knowledge (Smith and Plackner, 2010) on a series of items during the test. Yet, these indicators might not always serve best in identifying misfit within a CAT context. For instance, to detect misfit caused by test fraud, including item memorization, pre-existing item knowledge, or item parameter drift, alternative approaches to diagnosing misfit may be required. Alternatively, Walker and Engelhard (2016) proposed a two step-approach for person misfit detection that integrates person response functions (PRF, Trabin & Weiss, 1979) to person fit statistics. Their approach enables to further investigate reason and location of misfit. Another piece of graphical evidence could be grounded in the adaptive nature of CATs. As the CAT progresses to later stages, variability in ability estimates is expected to decrease. Plotting the ability estimates against the sequence of item administration and drawing a line through these estimates can offer further visual insight into person misfit. Ideally, in a typical CAT administration, the slope of this line should approach to zero, indicating stabilization in the ability estimation process. Otherwise, a deviation from this pattern would signal a person's misfit and warrants further investigation.

Overall, reviewing the available literature on person fit in CAT, it appears there remains significant room for research, particularly in light of recent advancements in CAT research, such as multistage testing. The points below highlight essential areas for further investigation and aims to offer a foundation for researchers interested in exploring this field more deeply:

- Specific Challenges in Measuring Affective Constructs with CATs: CAT applications of psychological constructs can yield unique challenges in person-fit analysis, such as biases linked to social desirability. Developing indices specially designed for the nature of affective CATs, which address the varied reasons for person misfit in assessing psychological constructs, could be a viable approach.

- Holistic Fit Indices for Multi-scale CAT: Utilizing CAT to evaluate individuals across a range of dimensions, from cognitive abilities and personality traits to specific skill sets, is known for its precision and efficiency on individual scales (Maurelli & Weiss, 1981). A composite fit index that considers the interrelationships and collective performance across scales could increase the CAT's effectiveness, ensuring a holistic assessment of person fit.

- Lastly, investigating person fit within multistage CAT applications can offer a promising avenue for research, especially in light of recent studies such as Sideridis, Ghamdi & Zamil (2023), which compare the effectiveness of multistage CAT and traditional CAT. Their findings highlight a notable divergence in theta scores for high-ability examinees within multistage CAT frameworks, despite generally supporting multistage CAT's role in enhancing measurement accuracy. This discrepancy highlights the necessity for further exploration into how different multistage CAT designs handle misfit detection, particularly in scenarios involving high and low-ability examinees.

References

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

Bradlow, E. T., & Weiss, R. E. (2001). Outlier measures and norming methods for computerized adaptive tests. Journal of Educational and Behavioral Statistics, 26(1), 85-104. https://doi.org/10.3102/10769986026001085

Chen, J., & Gupta, A. K. (2012). *Parametric statistical change point analysis: With applications to genetics, medicine, and finance* (2nd ed.). Springer.

Csörgő, M., Horváth, L., & Szyszkowicz, B. (1997). Integral tests for suprema of Kiefer processes with application. Statistics & Risk Modeling, 15(4), 365-378. https://doi.org/10.1524/strm.1997.15.4.365

Ferrando, P. J., & Anguiano-Carrasco, C. (2012). Response Certainly, Conscienciousness, and Self-concept Clarity as antecedents of Acquiescence: A prediction model. Anuario de Psicología, 42(1), 103-112. https://psycnet.apa.org/record/2014-14293-007

Hendrawan, I., Glas, C. A., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. Applied Psychological Measurement, 29(1), 26-44. https://doi.org/10.3390/educsci10110324

Maurelli, V. A., & Weiss, D. J. (1981). Factors Influencing the Psychometric Characteristics of an Adaptive Testing Strategy for Test Batteries. http://files.eric.ed.gov/fulltext/ED212676.pdf

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. Applied Psychological Measurement, 23(2), 147-160. https://doi.org/10.1177/01466219922031275

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

Meijer, R.R., van Krimpen-Stoop, E.M.L.A. (2009). Detecting Person Misfit in Adaptive Testing. In: van der Linden, W., Glas, C. (eds) Elements of Adaptive Testing. Statistics for Social and Behavioral Sciences. Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-8_16

Robin, F. (2002). Investigating the relationship between test response behavior, measurement and person fit. In annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sideridis, G., Ghamdi, H., & Zamil, O. (2023). Contrasting multistage and computer-based testing: score accuracy and aberrant responding. Frontiers in Psychology, 14, 1288177. https://doi.org/10.3389/fpsyg.2023.1288177

Sinharay, S. (2016). Person fit analysis in computerized adaptive testing using tests for a change point. Journal of Educational and Behavioral Statistics, 41(5), 521-549. https://doi.org/10.3102/1076998616658331

Smith, R. M., & Plackner, C. (2010). The family approach to assessing fit in Rasch measurement. In M. Garner, G. Engelhard Jr., W. Fisher Jr., & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 1, pp. 64–85). JAM Press.

Trabin, T. E., & Weiss, D. J. (1979). The person response curve: Fit of individuals to item characteristic curve models (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program. https://apps.dtic.mil/sti/tr/pdf/ADA080933.pdf

van Krimpen-Stoop, E.M.L.A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Kluwer-Nijhoff.

van Krimpen-Stoop, E.M.L.A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. Applied Psychological Measurement, 23, 327-345. https://psycnet.apa.org/doi/10.1177/01466219922031446

Walker, A. A., & Engelhard, G. (2016). Using person fit and person response functions to examine the validity of person scores in computer adaptive tests. In Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings (pp. 369-381). Springer Singapore. http://dx.doi.org/10.1007/978-981-10-1687-5

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 5-17

Investigation of Measurement Precision and Test Length in Computerized Adaptive Tests under Different Conditions

Hüseyin YILDIZ* Ceren TUNABOYLU DEMİR** Süleyman ÜLKÜ***
Gamze GİRAY**** Hülya KELECİOĞLU*****

Abstract

In this study, it is aimed to examine item exposure rate, content balancing, and ability estimation in terms of termination rules with regard to testing lengths and testing accuracy in computerized adaptive testing. In this context, EAP and MLE ability estimation methods were compared in terms of correlation, bias, RMSE, and test length. In the study EAP and MLE were compared with a total of 72 different conditions; including 1, 2, and 4 group content balancing patterns; 0.50, 0.75, and 1.00 exposure rates; 0.35 and 0.40 standard error-based and the termination rule based on the test length of 15 and 30. This research is Monte-Carlo simulation study, which was carried out in relational screening model of the quantitative research methods. The production and analysis of the data were performed in the Rstudio. As a result, the best performance in the measurement is a fixed test length of 30 items with 0.35 standard error; in group 1 pattern where the content balancing is not a group limitation; the exposure rate was displayed in the range of 0.75 and 1.00. Depending on the test length of ability estimation methods, scope balancing patterns and exposure rates, the number of items changes in the range of 22 and 25; Based on the termination rule, it was estimated that at least 0.40 standard errors with a standard error based on 39 items.

Keywords: computerized adaptive testing, contend balancing, exposure rate, simulation study

Introduction

With the developments of technology field, the need for the use of computerized adaptive testing (CAT) instead of the classical paper-pencil tests in the measurement and evaluation applications has increased, and the studies have become widespread. CAT is the form of creating tests, testing individuals and scoring individuals in the computer environment (Reckase, 2009). The most important feature that separates CAT from the paper-pencil tests is that how the test starts, continues and terminates may differentiate according to the individual. The individualization mentioned here works as a set of algorithms and rules.

Classical Test Theory (CTT) was used in the first examples of CAT applications (Betz & Weiss, 1973; Larkin & Weiss, 1974; Vale & Weiss, 1975). In CTT, test and item parameters may vary according to the ability level of the group. Due to its parameter invariance feature, Item Response Theory (IRT) eliminates this disadvantage of CTT. In IRT, item parameters do not change according to the ability

* Researcher, Australian Council for Educational Research (ACER), Methodology and Measurement Department, Melbourne-Australia, huseyin.yildiz@acer.org ORCID ID: 0000-0003-2387-263X
** Branch Manager, Republic of Türkiye Ministry of National Education, General Directorate of European Union and Foreing Relations, Ankara-Türkiye, cerentunaboylu@gmail.com, ORCID ID: 0000-0001-8090-8913
*** National Education Expert, Republic of Türkiye Ministry of National Education, General Directorate of Lifelong Learning, Ankara-Türkiye, ulkusuleyman@gmail.com, ORCID ID: 0000-0003-1965-0671
**** Phd. Candidate, Hacettepe University, Faculty of Education, Ankara-Türkiye, giraygamze@gmail.com, ORCID ID: 0000-0002-5795-4521
***** Prof.Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Yıldız, H., Tunaboylu Demir, C., Ülkü, S., Giray, G. & Kelecioğlu, H. (2024). Investigation of measurement precision and test length in computerized adaptive tests under different conditions, *15*(1), 5-17. https://doi.org/10.21031/epod.1068572

Received: 5.02.2022
Accepted: 1.06.2023

distribution of individuals who take the test. The predicted ability parameters do not change according to the properties of the items in the applied test (Hambleton, Swaminathan, & Rogers, 1991). IRT is a model that explains the relationship between an individual's ability level and his/her responses to the measured feature with a mathematical function (Embretson & Reise, 2000; Hambleton & Swaminathan, 1989). Although CAT applications are not dependent on IRT, the preference for IRT in CAT applications allows the results to be more effective (Weiss, 1983).

CAT starts with choosing items to start the test, estimating the ability of the test taker according to the responses given by them, and administering next item chosen based on the estimated ability level or stopping the test (Wainer, 2000). The process must be planned very well in a detailed manner to serve the purpose of the application.

The pre-condition of the CAT application is the creation of a large pool of test items. In order to achieve the advantages of CAT over paper-pencil tests, the item pool must contain high quality items in accordance with different ability levels (Flaugher, 2000). In the item pool used, sufficient number of items in accordance with each ability level must be present (Reckase, 1989). The CAT consists of four basic processing steps, including the starting of the test, item selection method, ability estimation and test termination rules (De Ayala & Koch, 1995; Reckase, 1989; Thompson & Weiss, 2011). The test begins with choosing the first item. The test can be initiated with the best distinctive or average difficulty item in accordance with the preliminary information about ability level of the individual. After the estimation of the individual's ability level according to the given response, the second item suitable for individual from the item pool is selected using different methods. A new ability estimation is performed according to the responses to the items. According to different termination criterion, the test is terminated, and the ability level of the individual is determined. As a result, the CAT application has a cycle in which the creation of the item pool, how to select the items from the item pool, how the scoring will be done, and in which situations the application will be stopped, are determined and implemented.

In CAT applications, the item to be answered by the individual is determined according to his/ her measured ability level. In this way, in test applications where maximum performance is measured, individuals with high ability levels take the more difficult items, while individuals with low ability levels take the easier items and a customized test form is formed for each individual. The basic rationale behind individualized tests is to avoid items that may be too difficult or too easy for the person taking the test and to choose the items that best suit the individual's ability. Individuals take items that provide more information for their own ability level, so that differences between individuals can be determined more clearly (Wang, 2013). The process results in shorter tests, as individuals receive items suitable for their ability level and do not waste their time dealing with more difficult or easier items for themselves (Wainer, 2000).

One of the most important advantages of CAT is its estimation individuals' abilities with a small number of items in comparison to the classical paper-pencil tests. Embretson and Reise (2000) stated that a well-patterned CAT application could reduce the test length up to 50% without significantly losing measurement accuracy. In classical paper-pencil tests, the measurement precision may vary according to the ability levels of the individuals in the group, while accurate measurements can be made according to the ability levels of the individuals in the CAT applications. However, there are problems such as the individual does not have a chance to return to the previous item in CAT applications, security violation problems caused by the disclosure of the item pool and the frequency of item use (Aybek et.al., 2014). However, in CAT applications, problems such as the individual's lack of a chance to return to the previous item, a security violation based on the disclosure of the item pool, and the frequency of item use may occur.

The item pools used in CAT applications include a large number of items. However, in some cases, the frequency of use is seen to be rather high for some of the items and for some items pretty low. When such cases are encountered, the possibility of recalling of items for the individuals can be considered high. In order to increase the utility level of the item pool, Item Bank Constraining methods have been developed. These methods are counted among the basic components of CAT applications since they offer solutions for the application problems (Davis, 2002; Boyd, 2003). These methods include Content Balancing and Item Exposure Rate.

Content Balancing

In the tests containing two or more contents, the items may vary with low and high ability levels in accordance with the content. Student group with any level of ability may be exposed to a restricted test. For example, Mathematics course, let's think about a test where four operations skills are included in a single scope. While the student group at the high ability level may only encounter with items for division skill, the student group at the low ability level may only encounter items for addition skills. In such a case, content balancing is needed.

The tests in which the content balancing is used are longer than the tests which is not. The ability, personality and preference measurements are relatively more homogeneous and one-dimensional in the content, so they do not need content balancing; however, content balancing is required for achievement tests (Weiss, 2004).

Content balancing methods can be evaluated into two categories those based on a methodological approach and approaches that select items by trying to meet a complex set of constraints (Demir, 2019). In the first approach, an item pool is divided into several sub-pools based on item attributes, and items are selected from sub-pools to meet predetermined content areas. On the other, it relies on a different approach, which makes item selection try to meet a complex set of constraints, and an item can contribute to satisfying more than one constraint at the same time.

There are Constrained CAT, Modified CCAT, and Modified Multinomial Model among the methods to ensure a fixed content balance (Lin, 2011). In addition, Weighted Deviations Model, Shadow Test Approach, Two-Phase Item Selection Procedure, Weighted Penalty Model, and Maximum Priority Index methods can be counted for large-scale applications (as cited in He, Diao, & Hauser, 2014).

In the literature, there are studies in which 2 to 6 content areas are determined and different content balancing methods are compared (Lin, 2011, Kingsbury & Zara, 2009; Kingsbury & Zara, 1980; Eggen & Netherland, 1999; Demir, 2019). These studies compared content balancing methods by keeping the number of content areas constant. In this research, using CAT, the cases where there is a different number (2, 4) of content areas and no content area were compared.

Item Exposure Rate

The use of item exposure rate is focused on protecting the integrity of the item pool and the confidentiality of the items in the item pool by blocking over-exposure to the same items (Davis & Dodd, 2005). With adaptive tests created from the same item pool, the same questions can be asked for multiple times and the individual can learn the correct answers. The most commonly used items lose their original psychometric properties by becoming popular. This situation causes the test validity to fall. The purpose of a good item exposure rate control method is to ensure the more balanced use of the item pool without reducing the measurement accuracy by defying this relationship (Pastor, Dodd & Chang, 2002). Item exposure rate control methods are used to balance the test security and measurement accuracy (Boyd, 2003; Boyd, Dodd & Fitzpatrick, 2013).

Item use frequency control methods prevent the disclosure of items by preventing excessive use of the same items, preserving the integrity of the substance pool (Davis & Dodd, 2005). It is possible to classify the item use frequency control methods into five categories. These are (1) randomization strategies, (2) conditional selection, (3) stratified strategies, (4) combined strategies, and (5) multiple-stage adaptive test designs (Lin, 2011). In this study, the frequency of use of the item was controlled by using the restricted maximum information strategy, which is one of the conditional selection methods. This method determines whether the item will be used when that item is selected by comparing it with the maximum value of the frequency of use parameter determined before the test.

Exposure rates were predetermined 0,10 and 0,20 (Chang & Ansley, 2003), 0,19 and 0,29 (Boyd et al., 2013), 0,30 (Pastor, Dodd & Chang), 0,40 (Revuelta & Ponsoda, 1998). In this research, two exposure

rates of 0,50 and 0,75 were studied due to the lack of a large item pool and the use of content balancing which is another restrictive method. To compare the effect of exposure rates, no exposure control condition was also added to the research.

When the literature is examined, there are studies in which many aspects of CAT (content balancing, item pool properties, test length, etc.) are compared under different conditions (Boyd, 2003; Eroğlu & Kelecioğlu, 2012; Demir, 2018; Aybek & Çıkrıkçı, 2018; Sulak & Kelecioğlu, 2019; Kara, 2019). It is considered that the research will contribute to the field in terms of examining the measurement accuracy and test length when ability estimation methods, content balancing approaches, item exposure rates and termination rules are changed in CAT applications. Based on the results of the research, it is evaluated that the research will contribute to the field of study by determining the conditions which provide calculations with minimum error and bias, and maximum correlation between true and estimated thetas.

Purpose of the Study

The main purpose of this study is to explain how bias, RMSE, correlation values between true and estimated thetas, and test length change according to different conditions of item exposure rate, content balancing, ability estimations methods and termination rules. Accordingly, the sub-problems of the study are given below.

a) How do bias, RMSE, correlation values between true and estimated thetas, and test length change according to different conditions of termination rules based on standard error (0,35 and 0,40), item exposure rate (0.50, 0.75, 1.00), content balancing (1 group, 2 groups and 4 groups) and ability estimation methods (Expected a Posteriori (EAP) and Maximum Likelihood (ML))?

b) How do bias, RMSE and correlation values between true and estimated thetas change according to different conditions of termination rules based on fixed length (15 and 30), item exposure rate (0.50, 0.75, 1.00), content balancing (1 group, 2 groups and 4 groups) and ability estimation methods (Expected a Posteriori (EAP) and Maximum Likelihood (ML))?

c) How do the average of bias, RMSE and correlation values between true and estimated thetas change in all conditions separately?

Simulation Methods

This study is a Monte Carlo simulation study that aims to reveal the relationship between various ability estimation methods, exposure rates, content balancing rules and termination rules in CAT applications. Collecting real data for research can be time-consuming and costly to collect. In addition, sometimes the use of real data may not be sufficient for the analyzes desired to be carried out in the research. In such cases, it may be more useful to generate the data. In the simulation study, the data is created by the researcher based on a model. Simulations have two major components. The first is a system that is of interest to the investigator, and the second is a model that represents the system. One advantage of simulation studies is that they allow researchers to compare estimated parameters against their respective true parameters, which are unknown for real data applications (Feinberg & Rubright, 2016; Wilcox, 1997). Also, simulation study is a quantitative relational research since it aims examining the relations between methods (Fraenkel & Wallen, 2006).

Data Generation

The data sets used in this study are produced by the help of the codes written by researchers in the R programming language. Fixed 200-item pools (Veerkamp & Berger, 1997) and 150 hypothetic participants (Guzmán & Conejero, 2004) are derived for each analysis. While producing ability parameters of individuals, standard normal distribution was used with mean of 0 and standard deviation of 1.

a Parameters of the items in the pools were obtained from the normal distribution of 0.8 mean and 0.1 standard deviation. b parameters used were obtained from uniform distribution in the range of (-3, + 3). Since the data generation is manufactured based on 2 parameter logistics model (2 PLM), c parameter (guessing parameters) is fixed 0.

Data Analysis

The CAT simulations were carried out with the "simulaterespondents" function in the "catR v.3.17" package of the programing language (Magis & Raiche, 2012). In this function, the ability parameters of individuals, item parameters, initial and termination rules are defined as compulsory arguments. In this study, the starting rule is fixed as an item that would generate the maximum information for a skill level to be chosen randomly in the range (-1.00,+1.00) for all analyses. Maximum Fischer Information, which is widely preferred in the literature (Choe, Kern, & Chang, 2017; Chen, Chao & Chen, 2019), was used as the item selection rule in all analyzes. This method is based on the principle of selecting the item that produces the highest information among the items in the estimated ability level after each response of the individual. The termination rules used are explained in the title of the simulation conditions because it is among the changed conditions. In this study, the correlation values between true theta scores and estimated theta scores were calculated with the Pearson correlation coefficient method, which is one of the parametric correlation methods.

Simulation Conditions

In this study, the ability estimation method, exposure rate, test termination rule and content balancing conditions were changed in CAT simulations. 3 different situations were used for content balancing. The first of these situations is not to use content balancing limitation, the second one is dividing item-pool into 2 content group, and the last situation is dividing the item-pool into 4 content group. In the conditions in which the item-pool had limitation of content balancing, analyzes were performed to be applied evenly for each group in terms of items applied.

Another condition that is changed in the study is the ability estimation method. To estimate ability Expected a Posteriori (EAP) and Maximum Likelihood (ML) ability estimation methods can be used. Maximum likelihood estimators are consistent, functions of sufficient statistics when sufficient statistics exist, efficient and asymptotically normally distributed (Hambleton & Swaminathan, 1985). The first of the methods used is the Maximum Likelihood Estimation (MLE) and the other is the Expected a Posteriori (EAP) method. MLE method is the most widely used method among the estimation methods based on the likelihood function, but it cannot give stable results when all answers are correct or incorrect. On the other hand, Bayesian approaches can make ability estimations for all response patterns (Embreston & Reise, 2000).

For CAT simulations, the "simulateRespondents" function allows us to determine both temporary ability estimates and final ability estimates. Within the scope of this study, the same method was used for temporary and final ability estimation.

Larger item pools are needed when content balancing and item exposure control methods are used to ensure content validity and test safety (Çoban, 2020). The item exposure rate restriction is used to allow an item in the item pool to be directed to a specified percentage of the group. In this study, 3 different exposure rate conditions were used for the restriction in question. In the first condition, the rate was accepted as 1.00. This rate means that the items are not brought to a restriction for the frequency of use. For 0.75 and 0.50 values used in other conditions, each item is allowed to be directed to the maximum of 75% and 50% of the groups respectively.

When to terminate the test is one of the important factors in estimating the ability level (Kezer, 2013). In this study, 4 different test termination rules are included. Two of these rules are based on the standard error limit of the ability estimation, while the other two conditions are the termination rules based on the fixed test length. While the fixed-length method is about the number of questions applied to the

individuals, the variable-length methods are related to the precision of the measurement. When the predetermined criteria are met in variable-length test termination methods, the individual's test is terminated. The minimum standard error method is the most widely used test termination method. According to this method, an individual's ability level depends on a certain standard error and if a certain measurement precision is reached, the test is terminated (Demir, 2018).As a standard error-based termination rule, 0.30 and 0.40 cutting scores were used. These values were frequently studied in the literature and were critical values in terms of test termination rules to obtain a measurement precision (Aybek & Çıkrıkçı, 2018; Sulak & Kelecioğlu, 2019; Yao, 2012). For termination rules based on fixed test length, 15 and 30 items are preferred. According to Stocking (1994), the item pool size should be at least 12 times the test length in CAT applications applied according to the fixed test length. Therefore, an item pool of 15 items was chosen to correspond to the item pool of 200 items, and an item pool of 30 items was chosen to disrupt this situation. 15 and 30 item values were preferred because they were found to be related to the test lengths obtained from the standard error-based termination rules of 0.35 and 0.40 in the preliminary analysis.

For 4 different changing conditions, 3x2x3x4= 72 simulation conditions were studied. We can not use the high number of replications (e.g. 100) used in simulation studies on different subjects. It is seen that 10-15 replications are made in similar simulation studies in which a large number of conditions are used. Basically, considering that everyone's CAT simulation is completed independently of each other and ability estimations are made separately, there is no difference between making 100 replications for 100 participants and 10 replications for 1000 participants. In addition, it took approximately 90 hours to complete 720 (72x10) simulations using a computer with high processing power, even under 10 replication conditions. So, for each condition, 720 different CAT analysis were performed in total with 10 replications (Gorin et.al., 2005; Kara, 2019).

Results

In this study, it is aimed to examine the measurement accuracy and test length of Computerized Adaptive Testing (CAT) when the ability estimation methods, content balancing patterns, exposure rates and termination rules are changed. In this context, EAP and MLE as an ability estimation, 1, 2, and 4 as content balancing pattern group patterns; 0.5, 0.75 and 1.00 as exposure rate, 0.35 and 0.40 standard error as termination rule and fixed length testing in the form of 15 and 30 items were used, and 72 different conditions were created and compared in terms of correlation, bias, RMSE and test length.

In the first stage, the analysis findings carried out according to 0.35 and 0.40 standard error-based termination rule are given in Table 1.

Table 1.

Correlation, Bias, RMSE and Test Length Results by Standard Error Termination Rule

Content Balancing	Estimation Method	Exposure Rate	Correlation		Bias		RMSE		Test Length	
			SE < 0,35	SE < 0,40	SE < 0,35	SE < 0,40	SE < 0,35	SE < 0,40	SE < 0,35	SE < 0,40
1 Group	EAP	0,50	0,867	0,612	0,006	0,037	0,496	0,785	35,823	9,565
		0,75	0,901	0,554	-0,026	-0,026	0,422	0,850	38,729	6,872
		1,00	0,907	0,695	0,003	0,016	0,409	0,701	38,414	13,692
	MLE	0,50	0,863	0,432	-0,028	-0,016	0,494	0,917	35,370	3,169
		0,75	0,894	0,603	-0,006	0,006	0,448	0,777	38,817	10,953
		1,00	0,921	0,514	0,028	-0,015	0,393	0,833	38,846	6,272

Table 1.

Correlation, Bias, RMSE and Test Length Results by Standard Error Termination Rule (Continued)

Content Balancing	Estimation Method	Exposure Rate	Correlation		Bias		RMSE		Test Length	
			SE 0,35	SE 0,40	SE 0,35	SE 0,40	SE 0,35	SE 0,40	SE 0,35	SE 0,40
2 Group	EAP	0,50	0,733	0,487	-0,058	-0,021	0,782	0,881	38,637	8,919
		0,75	0,749	0,590	-0,002	0,024	0,801	0,802	42,095	9,186
		1,00	0,871	0,570	0,012	0,038	0,508	0,780	39,365	11,136
	MLE	0,50	0,852	0,548	-0,023	0,016	0,528	0,850	37,140	8,522
		0,75	0,861	0,560	0,015	-0,035	0,492	0,828	38,951	7,355
		1,00	0,886	0,585	0,007	-0,056	0,481	0,786	39,580	10,300
4 Group	EAP	0,50	0,818	0,573	-0,007	-0,013	0,581	0,830	38,051	8,122
		0,75	0,865	0,610	-0,026	0,026	0,519	0,784	39,932	11,837
		1,00	0,850	0,665	0,028	0,042	0,538	0,734	40,042	14,427
	MLE	0,50	0,823	0,532	0,015	-0,012	0,577	0,864	37,755	8,270
		0,75	0,857	0,596	-0,032	-0,004	0,524	0,789	39,569	9,193
		1,00	0,842	0,551	0,050	-0,064	0,543	0,839	39,586	6,084

When the table 1 is examined, the lowest correlation value in all conditions of 0.35 standard error-based termination is 0.733, this value is the highest (0.695) in the analysis based on 0.40 standard errors. The bias values produced similar results in conditions where 0.35 and 0.40 standard error-based termination rules are used. For all circumstances, the bias values were approached to zero for 0.35 standard error when their absolute values are averaged. RMSE values range from 0.393 to 0.801 error-based termination rule for 0.35; and 0.701 and 0.864 for 0.40 standard error-based termination rule. In all conditions, RMSE values are estimated close to zero.

The test lengths range from 35.4 to 42.1 for 0.35 standard error-based termination criterion and from 3.2 to 14.4 for 0.35 standard error-based termination criterion. In this direction, it can be said that the condition of 0.35 standard error is used perform better than the condition that 0.40 standard error is used.

When content balancing conditions are compared under similar conditions, the correlation values are found to be higher in 1 group condition where there is no limitation in the item pool. The bias values have been met approximately similar to all conditions. RMSE values are generally relatively close to zero in 1 group condition. When the ability estimation is MLE, the deducted values are closer to each other in content balancing groups.

In the case of a comparison between EAP and MLE estimation methods, when 2 groups are used as content balancing condition, the MLE method was found to have higher correlation values, similar bias value, RMSE values closer to zero and relatively shorter test length. In cases where 1 and 4 groups are used as the content balancing condition, the values obtained in the EAP and MLE methods are similar.

When the values obtained according to exposure rates are examined, the correlation values are seen to reduce as the exposure rates decrease. The bias values are mostly higher in the EAP method for 0.75 exposure rate, while the MLE method is higher in 0.5 and 1 exposure rates. RMSE values are relatively estimated in both MLE and EAP methods as closer to each other and zero.

In the second stage, 15 and 30-item fixed test length is used according to the termination rules, the analysis findings are given in Table 2.

Table 2.

Correlation, Bias, and RMSE Results by Fixed Test Length Termination Rule

Content Balancing	Estimation Method	Exposure Rate	Correlation		Bias		RMSE	
			15	30	15	30	15	30
1 Group	EAP	0,50	0,853	0,904	0,031	0,002	0,534	0,438
		0,75	0,872	0,895	-0,034	-0,003	0,509	0,436
		1,00	0,850	0,900	-0,009	0,015	0,532	0,432
	MLE	0,50	0,836	0,905	0,019	0,001	0,548	0,424
		0,75	0,838	0,908	-0,015	0,011	0,514	0,415
		1,00	0,857	0,892	0,019	-0,004	0,523	0,464
2 Group	EAP	0,50	0,823	0,838	0,018	-0,009	0,577	0,555
		0,75	0,790	0,798	-0,005	-0,011	0,608	0,645
		1,00	0,796	0,789	-0,022	-0,043	0,621	0,649
	MLE	0,50	0,770	0,823	-0,028	-0,005	0,633	0,611
		0,75	0,828	0,908	0,003	0,011	0,548	0,419
		1,00	0,859	0,859	0,003	0,019	0,517	0,507
4 Group	EAP	0,50	0,866	0,909	0,010	0,017	0,512	0,411
		0,75	0,837	0,869	-0,007	0,015	0,540	0,486
		1,00	0,858	0,894	-0,027	0,021	0,521	0,438
	MLE	0,50	0,839	0,889	-0,010	0,020	0,563	0,455
		0,75	0,840	0,880	0,016	-0,020	0,545	0,463
		1,00	0,853	0,899	0,044	0,027	0,516	0,435

When Table 2 is examined, except for the content balancing 2 group condition, the estimation method EAP and exposure rate 1 condition, in all conditions correlation values based on the termination rule with fixed test length are higher in 30-item constant testing lengths. According to test lengths, bias values did not show a specific pattern according to the content balancing, estimation method and exposure rate. The average of the absolute values of the bias values for all conditions were found closer to zero for 30-item testing. Except for the content balancing 2 group condition, the estimation method EAP and exposure rate 0.75 and 1 conditions; RMSE values were estimated to be smaller for 30-item testing. In this respect, it can be said that the 30-item condition based on the fixed test length is better performed than 15-item condition.

Under similar conditions, when content balancing conditions are compared, the correlation values were higher in the case of 4 group limitations in the item pool. The bias values were predicted as similar. RMSE values are relatively close to zero in conditions of the 1 group, 4 groups and 2 group limits, respectively. The predicted values are closer to each other in content balancing groups when the ability estimation is MLE.

When EAP and MLE estimation methods are compared, the correlation values, RMSE and bias values are found to have no significant differences. When the values obtained according to exposure rates are examined, the correlation and bias values are generally reduced as the exposure rate decreases in the MLE method; however, RMSE and test length increase.

In order to facilitate the comments and comparisons obtained according to all conditions, the values obtained by the averages of all other conditions are given in Table 3. When the test length averages are taken, the fixed test lengths are not included in the mean.

Table 3.

Correlation, Bias, RMSE, and Test Length Averages by Simulation Conditions

	Conditions	Correlation	Bias	RMSE	Test Lenght
Termination Rule	0,35	0,853	0,021	0,530	38,7
	0,40	0,571	0,026	0,813	9,1
	15	0,837	0,018	0,548	-
	30	0,875	0,014	0,482	-
Estimation Method	EAP	0,785	0,020	0,601	24,7
	MLE	0,783	0,019	0,585	23,1
Content Balancing	1	0,803	0,016	0,554	23,0
	2	0,757	0,020	0,642	24,3
	4	0,792	0,023	0,584	24,4
Exposure Rate	0,50	0,766	0,018	0,619	22,4
	0,75	0,788	0,016	0,590	24,5
	1,00	0,798	0,026	0,571	24,8

When Table 3 is examined, 0.853 in the standard error-based termination rule of 0.853, in 0.40 standard error-based termination rule 0.571; in 15-item fixed test length 0.837, and in 30-item fixed test length 0.875 average correlation value was obtained. The highest correlation value was obtained in the condition of 30-item constant test length. In the case of 30-item fixed test length, the bias value was 0.014 and the RMSE value was 0.482, and these values were found to be closer zero in all other conditions.

The average test length was estimated as 38.7 in standard error-based termination rule as 0.35, and 9.1 in standard error-based termination rule in 0.40.

When the values are examined according to the estimation methods, it is estimated that the average correlation value in EAP condition is 0.785 while it is 0.783 in MLE. These values are quite close; and it is similar for the bias, RMSE and the test length. The average test length is estimated in EAP condition as 24.7, and 23.1 in MLE condition.

For content balancing 1 group condition, the average correlation value is 0.803, 0.757 in 2 group condition and 0. 792 in 4 group condition. The highest mean correlation value is in content balancing 1 group condition and the value of bias (0.016) and RMSE (0.554) is closest to zero (0.016). The average test length was 23.0 in content balancing group 1, 24.3 in 2 group condition, and 24.3 in t4 group condition.

The average correlation values according to exposure rate are 0.766 in 0.50, 0.75 in 0.788 and 1 in 0.798 in 1. The highest average bias value (0.016) was obtained in exposure rate 0.75, while the average bias value (0.016) is the smallest in exposure rate 0.75. RMSE is predominantly predicted in exposure rate 1 (0.571) closer to zero. The average test length was estimated 22.4 in exposure rate 0.50, 24.5in condition 0.75; 24.8 in condition 1.

Discussion and Conclusion

In general, when the results were examined, a standard error rule of 0.35 as a stop rule has performed in terms of better in terms of RMSE, correlation, and bias. Similarly, in previously research where they compared .30 and .40 standard error-based termination rule, Özbaşı and Demirtaşlı (2015) determined .30 standard error-based termination rule performed better than .40. 15 and 30-items test length conditions based on fixed test length are added to standard error-based comparisons, 30 items test performed better in terms of RMSE, correlation and bias among the four different termination rules. These findings are supported by the studies in the literature (Eroğlu & Kelecioğlu 2015; Lee, 2014; Calender, 2011; Babcock & Weiss, 2012).

When the content balancing methods were compared within themselves, the highest correlation was observed when the number of groups was 1 and low when the number of groups was 4. It was evaluated that there was no interpretable relationship between the number of groups and the correlation, since the values were not ordered depending on the number of groups and the difference between the correlation values was small. Leung, Chang, and Hau (2003) conclude that content balancing does not have a systematic effect on the measurement accuracy is in parallel with the findings of this study. In addition to this, in the study mentioned, group 1 pattern, where there were not coverage balancing limitation, performed well similarly. The reason why the content balancing influences the measurement accuracy is the restrictions it brings to the item bank. It can be interpreted that if the item pool is large enough and a parameters are or high or the number of content groups are few; content balancing will not affect measurement accuracy significantly.

A noticeable difference was not detected between EAP and MLE ability estimation methods. Similarly, in the studies carried out by Kezer (2014), Malak and Kelecioğlu (2019), they have not found a difference between EAP and MLE approaches. In addition, the analysis where EAP method used took much time. It can be said that the use of the MLE method may be preferred in terms of the economic use of time.

In the exposure rate, it can be said that the 1.00 condition is better performed than 0.75 and 0.50 conditions. However, in terms of the item security, the values obtained for the exposure rate at 1.00 and 0.75 are close, a value between 0.75 and 1.00 can be chosen for the exposure rate. In the case of exposure rate conditions similar to content balancing conditions, the differences between the correlation values obtained by 1.00, 0.75 and 0.50 were small. The exposure rate is also related to the restricting the use of item bank such as content balancing. The fact that the difference between the conditions are small is due to a sufficiently large item pool, it is thought that other items selected from the wide pool can make similar estimations at certain points.

As a result, according to the results of the simulation, 0.35 standard error based 30 items fixed length based termination rules; 1 group content balancing and between 0.75 and 1.00 exposure rate conditions were seen to perform better in terms of RMSE, correlation and bias. In addition, in terms of test length, ability estimation methods, content balancing patterns and exposure rates are estimated approximately 22 to 25, while the standard error-based termination rules are predicted to be 39 for 0.35 standard error and 9 for 0.40 standard error.

The average correlation value obtained with a 15-item fixed termination rule is 0.837; This value is 0.875 for 30-item conditions. It was observed that there is only a 0.03 increase in the correlation value in terms of test length between these two finishing rules. In this respect, 15-item termination condition is considered more efficient.

In order to better observe the effects of content balancing and exposure rates, especially on the measurement, it is thought that similar studies can be carried out with smaller item pools, with a greater number of content groups, or lower exposure conditions. It is also thought that the content balancing and exposure ratio variables can interact with the item selection methods. Therefore, a similar study can also be performed by changing item selection methods.

In addition, because the low values of the standard error increase the measurement accuracy, practitioners can determine the appropriate standard error according to the vitality of the test. If the content balancing method is used, larger item pool is needed in each content area. Since there is no significant difference between the MLE and EAP methods in terms of measurement accuracy, the MLE method can provide to practitioners an advantage in terms of analysis time. In the large of item pool, the item exposure rate did not differ much between 1 and 0.75. Practitioners should use a large item pool if they want to use a item exposure rate.

Declarations

Author Contribution: Hüseyin YILDIZ: data analysis, conceptualization, investigation, methodology, visualization, writing - review & editing. Ceren TUNABOYLU DEMİR: conceptualization, investigation, writing - review & editing. Süleyman ÜLKÜ: conceptualization, investigation, writing - review & editing. Gamze GİRAY: conceptualization, investigation, writing - review & editing. Hülya KELECİOĞLU: writing-review & editing, supervision

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 the data in this study were generated by a computer program.

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

References

Chen, L-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model.* [Doctoral dissertation, The University of Texas]. UT Electronic Theses and Dissertations. https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-12-344

Choi, Y. J., & Asilkalkan, A. (2019). R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 168-175. https://doi.org/10.1080/15366367.2019.1586404

Aybek, E., & Çıkrıkçı, R. (2018). Kendini Değerlendirme Envanteri'nin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliği [Applicability of the self assessment inventory as a computerized adaptive test] . Türk Psikolojik Danışma ve Rehberlik Dergisi, 117-141. https://dergipark.org.tr/en/pub/tpdrd/issue/40299/481364

Aybek, E. C., Şahin, D. M., Eriş, H. M., Şimşek, A. S., & Köse, M. (2014). Kağıt-kalem ve bilgisayar formunda uygulanan testlerde öğrenci başarısının karşılaştırıldığı çalışmaların meta-analizi [Meta-analysis of comparative studies of student achievement on paper-pencil and computer-based test]. Asya Öğretim Dergisi, 2(2), 18-26. https://dergipark.org.tr/en/pub/aji/issue/1539/18831

Babcock, B. & Weiss, D.J. (2012). Termination criteria in Computerized Adaptive Tests: do variable-length CAT's provide efficient and effective measurement? International Association for Computerized Adaptive Testing, 1, 1-18. https://doi.org/10.7333/1212-0101001

Boyd, M. A. (2003). Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems. Unpublished Doctoral Thesis, The University of Texas, Austin.

Boyd, A. M., Dodd, B., & Fitzpatrick, S. (2013). A Comparison of Exposure Control Procedures in CAT Systems Based on Different Measurement Models for Testlets. Applied Measurement in Education, 113–135. https://doi.org/10.1080/08957347.2013.765434

Chen, J.-H., Chao, H.-Y., & Chen, S.-Y. (2019). A Dynamic Stratification Method for Improving Trait Estimation in Computerized Adaptive Testing Under Item Exposure Control. Applied Psychological Measurement, 1-15. https://doi.org/10.1177/0146621619843820

Davis, L. L. (2002). Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Polytomously Scored Items. Unpublished Doctoral Thesis, The University of Texas, Austin.

Davis, L. L., & Dodd, B. G. (2008). Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Partial Credit Model. Pearson Educational Measurement, , 9(1), 1. https://doi.org/10.1177/0146621604264133

Demir, S. (2018). Çok kategorili bireyselleştirilmiş bilgisayarlı test uygulamalarının farklı madde seçim yöntemlerinde sonlandırma kuralları açısından incelenmesi [Investigation of Different Item Selection Methods in Terms of Stopping Rules in Polytomous Computerized Adaptive Testing]. [Unpublished Doctoral Thesis]. Hacettepe University, Ankara.

Demir, S. (2019). Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinde Sınıflama Doğruluğunun İncelenmesi [Investigation of Classification Accuracy at Computerized Adaptive Classification Tests]. [Unpublished Doctoral Thesis]. Hacettepe University, Ankara.

Choe, E., Kern, J., & Chang, H.-H. (2017). Optimizing the Use of Response Times for Item Selection in Computerized Adaptive Testing. Journal of Educational and Behavioral Statistics, 1-24. https://doi.org/10.3102/1076998617723642

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Taylor & Francis.

Eroğlu, M. G., &Kelecioğlu, H. (2012). Bireyselleştirilmiş bilgisayarli test uygulamalarinda farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması [Comparison of different test termination rules in terms of measurement precision and test length in computerized adaptive testing]. Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 28(1), 31-52. https://doi.org/10.19171/uuefd.87973

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. (2000). Computerized Adaptive Testing: A Primer Second Edition (s. 37-59). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9781410605931

Fraenkel, J., & Wallen, N. (2011). How to design and evaluate research in education (6th ed.). McGraw-Hill, Inc.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. Applied Psychological Measurement, 29(6), 433-456. https://doi.org/10.1177/0146621605280072

Guzmán, E., & Conejo, R. (2004, August). A model for student knowledge diagnosis through adaptive testing. In International Conference on Intelligent Tutoring Systems (pp. 12-21). Springer. https://doi.org/10.1007/978-3-540-30139-4_2

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory. Springer.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. Educational and Psychological Measurement, 74(4), 677-696. https://doi.org/10.1177/0013164413517503

Kalender, İ. (2011). Effects of Different Computerized Adaptive Testing Strategied on Recovery of Abilitiy. Unpublished Doctoral Thesis, Middle East Technical University, Ankara.

Kara, B. E. (2019). Computer adaptive testing simulations in R. International Journal of Assessment Tools in Education, 6(5), 44-56. https://doi.org/10.21449/ijate.621157

Kezer, F. (2013). Bilgisayar Ortamında Bireye Uyarlanmış Test Stratejilerinin Karşılaştırılması [Comparision of The Computerized Adaptive Testing Strategies].[Unpublished Doctoral Thesis]. Ankara University, Ankara.

Lee, M. (2014). Application of Higher-Order IRT Models And Hierarchical IRT Models To Computerized Adaptive Testing. Unpublished Doctoral Thesis, University of California, Los Angeles.

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. Applied Psychological Measurement 71(1), 20-36. https://doi.org/10.1177/0013164410387336

Magis, D., & Raiche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R package catR. Journal of Statistical Software, 1-31. https://doi.org/10.18637/jss.v048.i08

Özbaşı, D., & Demirtaşlı, N. (2015). Bilgisayar okuryazarlığı testinin bilgisayar ortamında bireye uyarlanmış test olarak geliştirilmesi [Development of computer literacy test as computerized adaptive testing]. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6(2), 218-237. https://doi.org/10.21031/epod.79491

Pastor, D. A., Dodd, B. G., & Chang, H.-H. (2002). A Comparison of Item Selection Techniques and Exposure Control Mechanisms in CATs Using the Generalized Partial Credit Model. Applied Psychological Measurement , 147-163. https://doi.org/10.1177/01421602026002003

Reckase, M. D. (2009). Designing item pools to optimize the functioning of a computerized adaptive test. Psychological Test and Assessment Modeling, 127-141.

Sulak, S., & Kelecioğlu, H. (2019). Investigation of Item Selection Methods According to Test Termination Rules in CAT Applications. Journal of Measurement and Evaluation in Education and Psychology, 315-326. https://doi.org/10.21031/epod.530528

Thompson, N. A., & Weiss, D. J. (2011). A Framework for the Development of Computerized Adaptive Tests. Practical Assessment, Research & Evaluation., 1- 9. https://scholarworks.umass.edu/pare/vol16/iss1/1/

Yao, L. (2013). Comparing the Performance of Five Multidimensional CAT Selection Procedures With Different Stopping Rules. Applied Psychological Measurement, 3-23. https://doi.org/10.1177/0146621612455687

Weiss, D. J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. Measurement and Evaluation in Counseling and Development, 70-84. https://doi.org/10.1080/07481756.2004.11909751

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), Computerized adaptive testing: A primer. Lawrence Erlbaum Associates.

Wilcox, R. R. (1997). Simulation as a research technique. In J. P. Reeves (Ed.), Educational research, methodology, and measurement: An international handbook (pp. 150–153). Pergamon.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics, 22(2), 203–226. https://doi.org/10.3102/10769986022002203

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 18-34

Measurement Invariance of Achievement Motives Model: PISA 2018 Turkey Sample

Münevver BAŞMAN*

Abstract

This study aims to find out whether the achievement motives model, constructed by attitudes towards competition, motivation to master tasks, and fear of failure scales, has measurement invariance in the PISA 2018 student questionnaire concerning gender and school type in Turkey sample, containing 6442 students. According to the results, the model's fit levels with the data were within acceptable levels across gender groups and school-type groups. Then, the measurement invariance across gender and school type was tested by multigroup confirmatory factor analysis including a sequence of tests of four nested hierarchical models which are configural, metric, scalar, and strict invariance. The fit indices of models and the differences of indice values between models were examined to decide whether measurement invariance is established. It is found that the full measurement invariance holds according to gender and school type since the values of the indices for each invariance step are acceptable. It means that it will be appropriate and meaningful to compare the students based on the scores obtained from the achievement motives model.

Keywords: achievement motives, gender, measurement invariance, PISA, school type

Introduction

International assessments allow countries to observe their successes and shortcomings as well as their situation compared with other countries. One of these international assessment studies conducted in this direction is the Programme for International Student Assessment (PISA) and provides important data for educational and social research. PISA reveals the school success of students and handles the factors affecting their performance, as well as allowing comparisons between countries.

In PISA administrations, students are assessed every three years in three subjects: reading, mathematics, and science. Every three years, only one of these areas constitutes the main subject of the application. PISA started with reading literacy as the major domain in 2000 and then continued with the main fields of mathematics and science, respectively. This process has continued in this order until now. In these administrations, cognitive tests are applied to see the extent to which 15-year-old students have the knowledge and skills necessary for participation in societies, while questionnaires are implemented to assess student background factors, school-level factors, and non-cognitive and metacognitive factors. As in previous cycles of PISA, PISA 2018 student questionnaires dealt with non-cognitive and metacognitive variables related to the main subject (reading-related outcomes). In addition to this, it is concerned with non-cognitive variables (dealing with general topics rather than domain-specific topics) called dispositional variables and school-focused variables (learning beliefs and attitudes towards school and achievement goals).

Dispositional Variables in PISA Questionnaires

Dispositional variables are the personality-based contexts that include students' approaches to learning or their avoidance, such as the achievement motives of competitiveness, fear of failure, and work

* Assist. Prof. Dr. Marmara University, Faculty of Education, İstanbul-Türkiye, munevver.rock@gmail.com, ORCID ID: 0000-0003-3572-7982

To cite this article:
Başman,M. (2024). Measurement invariance of achievement motives model: PISA 2018 Turkey sample, *15*(1), 18-34. https://doi.org/10.21031/epod.1302574

Received: 25.05.2023
Accepted: 13.01.2024

mastery; subjective well-being; perseverance; incremental mind-set; and information and communication technology motivation and practices. It is the result of lifelong socialization by parents, teachers, coaches, and one's cultural environment, and shows how behavior gets stronger over time. These variables are important since they are one of the best predictors of achievement and domain-specific outcomes (Organization for Economic Co-operation and Development [OECD], 2019a).

One of the dispositional variables determined in PISA 2018 is achievement motives constructed with work mastery, competitiveness, and fear of failure variables. Henry Murray (1938; as cited in Hangen & Elliot, 2016) introduces the achievement motives and Atkinson (1957) presents a model in which achievement motives are the figures that motivate people to be successful and avoid being unsuccessful in some standards of excellence in certain conditions. Two concepts are mentioned here: the need for achievement and the fear of failure. The need for achievement is introduced with three factors: mastery ("preference for challenging, difficult tasks"), work ("enjoyment of working hard"), and competitiveness ("liking for interpersonal competition and the desire to better others") in the Work and Family Orientation Questionnaire (WOFO) developed by Helmreich and Spence (1978) (Helmreich et al., 1980, p. 4). Then, mastery and work factors are combined as a work mastery motive because mastery and work factors are highly correlated and share important content. This creates a two-dimensional model, work mastery and competitiveness, of the need for achievement (Spence & Helmreich, 1983, as cited in Hangen & Elliot, 2016).

The other factor of the achievement motive is competitiveness. Franken and Brown (1995) state that individuals want to be competitive for different reasons and try to identify these reasons in their study. They develop a scale with five factors ("desire to win, satisfaction that comes from improving one's performance, motivation to put forth effort in competitive situations, satisfaction that comes from performing well, preference for difficult tasks") (Franken & Brown, 1995, p. 178). The first three factors are associated with competitiveness and the last two factors are about work mastery motives, which are the factors in the Questionnaire of Spence and Helmreich mentioned before.

The last variable of the achievement motives is fear of failure. It is defined as "disposition to avoid failure and/or a capacity for experiencing shame and humiliation as a consequence of failure" by Atkinson (1957, p. 360). Shame seems to be an emotional consequence of failure, which is highly disturbing to individuals with a high fear of failure, and has been shown to be associated with avoidance and withdrawal tendencies (Mascolo & Fischer, 1995). In other words, it is a tendency that focuses on avoiding the consequences of failure, unlike the need for achievement (Hangen & Elliot, 2016). The consequences of failure are feared rather than the failure itself (Birney et al., 1969, as cited in Conroy, 2003).

Competitiveness, work mastery, and fear of failure are also defined in the PISA 2018 reports. Competitiveness is stated as a desire to be superior to others. Work mastery is described as a desire to work hard to complete tasks. Fear of failure is expressed as a tendency to avoid potential errors and failures in order to protect themselves. Other PISA cycles assess similar factors but these factors are reviewed and reconstructed in PISA 2018 as the factors of achievement motives. For example, test anxiety was used in the previous PISA cycles, but fear of failure is used instead of test anxiety in PISA 2018. It is stated that fear of failure is a more general tendency to avoid potential mistakes and failures because they are experienced as embarrassing, and this can predict cognitive achievement in real-life situations more than test anxiety (OECD, 2019a).

The Measurement Invariance

The achievement motives can affect students' achievements directly or indirectly. However, when the results obtained from or related to these variables are compared between groups, it is not correct to attribute the differences only to the characteristics of the groups, because these differences between the groups may be due to the measurement tool rather than the characteristics of the groups. It is not certain whether any difference between the groups is because of a true difference or psychometric differences (Cheung & Rensvold, 2002). Differences in scores may be due to many confounding variables, such as familiarity with item response formats, test adaptation, and many other socio-cultural factors. Groups can only be compared when scale scores from different groups measure the same factor of interest on the same metric. Only then can score differences between groups be truly represented and meaningful. Therefore, evidence should be presented to make a factor comparison

across groups (Wu et al., 2007). One of these pieces of evidence is measurement invariance evidence. Drasgow and Kanfer (1985) state that measurement invariance is established when the relationship between observed scores and latent factors is the same across groups and when individuals from different groups having the same scores on the latent factor have the same observed scores. In other words, it means that the probability of an individual's observed score being independent of group membership depends on the true score (Wu et al., 2007).

There are various methods for examining measurement invariance. Khorramdel et al. (2020) indicate that some researchers interested in cross-cultural tradition have given their attention to measurement invariance in non-cognitive measures using the latent variable framework and multigroup confirmatory factor analysis (MGCFA). MGCFA, introduced by Jöreskog (1971), is one of the methods of structural equation models used to determine the measurement invariance. MGCFA examines a large number of issues through a single procedure rather than through many separate procedures. Structural Equation Modeling (SEM) provides direct measurement of how much a measurement model is improved or impaired by various intergroup constraints; this offers a clear advantage over other techniques currently in use (Cheung & Rensvold, 2002). On the other hand, MGCFA has disadvantages and limitations in testing measurement invariance when the number of groups and sample size in the data are large (Ding et al. 2023). Measurement invariance with MGCFA is examined by testing four nested hierarchical models or hypotheses, which are: configural invariance, metric (weak) invariance, scalar (strong) invariance, and strict (residual) invariance (Meredith, 1993; Steenkamp et al., 1998; Vandenberg & Lance, 2000).

Configural invariance is the basic form and the first step of invariance. It is tested whether factors have the same pattern of free and fixed loadings across groups and whether individuals in different groups use the same conceptual framework when answering the scale items (Cheung & Rensvold, 2002; Khojasteh, 2012; Wu et al., 2007; Vandenberg & Lance, 2000). Metric invariance is the equality test for scaling units across groups. It determines whether the item loadings on the factors are the same across groups (Khorramdel et al., 2020). Factor loadings are regression slopes that connect the observed variables to the latent variables of interest and thus represent the expected amount of change in the observed variable for one unit of change in the latent variable (Wu et al., 2007). Scalar invariance is the equality test of the intercepts of the regression equations of the observed scores on the latent variables across groups (Khademi, 2020; Schmitt & Kuljanin, 2008). It is tested whether the mean differences in the observed scores are attributed to the mean differences of the latent variables (Finch & French, 2015; Steinmetz et al., 2009; Tucker et al., 2006). Strict invariance is the equality test of unique variances across groups (Khademi, 2020; Vandenberg & Lance, 2000). It is tested whether the mean or covariance differences in the observed scores are attributed to the mean or covariance differences in the latent variables (Gregorich, 2006; Meade et al., 2006).

The measurement invariance of the questionnaires related to the achievement motives and structures in the PISA application was determined in order to determine the usability of the questionnaires in Turkey. The measurement invariance of various scales in these questionnaires was examined according to some variables such as gender, school type, statistical region, socioeconomic status, and countries. In this study, the measurement invariance of the relevant model is handled according to gender and school type. In PISA applications, the relationship between various information obtained from students through questionnaires and students' literacy performance is examined. In the PISA final reports, the success differences of students in school types and different gender groups and the factors affecting success are discussed in detail (Education Reform Initiative-ERG, 2009; OECD, 2019b). It is a common finding of international and national studies that academic achievement differences between gender and school types have existed for a long time in Turkey (Berberoğlu & Kalender, 2005; Suna et al., 2020). In comparisons of questionnaires and tests by gender and school type, it is assumed that the measurements are equally valid in different groups, and measurement invariance can be ignored. The studies conducted in Turkey in the last 10 years examining the measurement invariance of the relevant structure according to gender and school type are given below.

Researchers have examined the invariance of the scales or models in the PISA survey according to gender, school type, countries, statistical region, socioeconomic status, and years (Ardıç & Gelbal , 2017; Başusta & Gelbal, 2015; Demir, 2016; Gülleroğlu, 2017; Güngör & Atalay Kabasakal, 2020; İmrol, 2017; Kıbrıslıoğlu, 2015; Kıbrıslıoğlu Uysal & Akın Arıkan, 2018; Uyar & Doğan, 2014; Uyar

& Kaya Uyanık, 2019). It has been observed that some models provide full measurement invariance (measurement invariance in all four steps is supported) according to the relevant variables, while others do not. For example, while Başusta and Gelbal (2015), Kıbrıslıoğlu (2015), Gülleroğlu (2017), Kıbrıslıoğlu Uysal and Akın Arıkan (2018), and Güngör and Kabasakal (2020) found full measurement invariance according to gender in the models established in their studies, Demir (2016), Ardıç and Gelbal (2017), and Uyar and Kaya Uyanık (2019) show that full measurement invariance has not been established according to gender.

Examining research conducted outside of Turkey, some studies (Adsul & Kamble, 2008; Awan et al., 2011; Nien & Duda, 2008; Shekhar & Devi, 2012; Tang & Lu, 2013) have demonstrated full measurement invariance across gender, while others (Freund et al., 2011; Karaman & Smith, 2019) have not. There are also studies examining differences in attitudes towards competition, motivation to master tasks, and fear of failure according to gender, and they found the full measurement invariance of the scales considering gender because the results of group differences obtained without measurement invariance are questionable (De Paola et al., 2015; Eber et al., 2021; Givord, 2020; OECD, 2019b; Severiens & ten Dam, 1998).

Whether it is PISA applications or other international applications, the results of these applications guide the development of education policies. In order for the results of the applications to be meaningful and valid, the measurement invariance of the measurement tools (achievement tests and questionnaires) used in the research should be ensured between subgroups such as gender, socio-economic level, school type, and culture; otherwise the comparisons will not be meaningful and valid (Vandenberg & Lance, 2000). It was observed that sufficient measurement invariance studies were not conducted on the student questionnaires of the PISA 2018 application. For this reason, examining the measurement invariance for the achievement motives model used in the PISA 2018 application will provide evidence for the validity of the model and determining whether the group comparisons are meaningful according to the scores obtained will contribute to a more accurate interpretation of the results. Thus, it is thought that examining the measurement invariance of the model, which has not yet been made in the literature, will fill the gap in the field. In Turkey, the main subject of PISA applications can be examined in terms of affective variables, and affective variables can be handled in terms of demographic variables such as gender and school type. Achievement differences between school types and gender in Turkey can be relatively high (Berberoğlu & Kalender, 2005; Suna et al., 2020). Examining the measurement invariance of the achievement motivation model in terms of gender and school type is important for Turkey in the context of equality in education. For these reasons, the aim of this study is to examine the measurement invariance of the achievement motives model constructed by attitudes towards competition, motivation to master tasks, and fear of failure scales in the PISA 2018 student questionnaire with regard to gender and school type in the Turkey sample. Answers were sought for the following questions in the study:

(1) What are the levels of fit of the achievement motives model with the data obtained from the whole group, gender, and school type subgroups?

(2) Does the achievement motives model hold measurement invariance across gender and school type subgroups?

Methods

Research Design

In this study, it is examined whether measurement invariance of the achievement motives model, including attitudes towards competition, motivation to master tasks and fear of failure scales, is held across gender and school type in the PISA 2018 application in the Turkey sample. This study is descriptive research and aims to determine an existing situation concerning the psychometric characteristics of the measurements obtained from the scales (Fraenkel et al., 2012; Karasar, 2019).

Population and Sample

In the PISA 2018 application, 38 OECD member countries and 41 non-member countries participated. There are 600,000 students, representing about 32 million in total (OECD, 2019b). In this research, the measurement invariance of the achievement model is examined in the Turkey sample. Turkey participated in the PISA 2018 application with 6890 students from 186 schools, representing approximately 884,971 students at the age of 15. Schools in determining the Turkey sample of the PISA 2018 application school type, Regional Units for Statistics Classification Level 1, administrative

form of the school, location of the school, and gender distribution layers were used. After the schools were determined, the students who would participate in the application at the selected schools were randomly selected (MEB, 2019). The Turkish sample consists of 6442 students. Table 1 presents the distribution of the students in the study groups according to their genders and school types.

Table 1.
Distribution of the Students in Study Group according to Gender and School Types

School	Gender				Total	
	Female		Male			
	n	%	*n*	%	*n*	%
Anatolian High School	1456	51.2	1386	48.8	2842	44.1
Vocational and Technical Anatolian High School	881	44.0	1122	56.0	2003	31.1
Anatolian Imam and Preacher High School	469	54.1	398	45.9	867	13.5
Science, Social Sciences, Multi-Programme Anatolian, Anatolian Sport/Anatolian Fine Arts High School	417	57.1	313	42.9	730	11.3
Total	3223	50.0	3219	50.0	6442	100

This study is carried out with 3223 (about 50%) female and 3219 (about 50%) male students. The schools attended by these students are Anatolian High School (44.1%), Vocational and Technical Anatolian High School (31.1%), Anatolian Imam and Preacher High School (13.5%), and Science, Social Sciences, Multi-Programme Anatolian, and Anatolian Sport/Anatolian Fine Arts High School (11.3%).

Data Collection

Data obtained from the PISA 2018 student questionnaire is used in this study. The data file for PISA 2018 can be found at the OECD PISA website, https://www.oecd.org/pisa/data/2018database/. Within the scope of the research, achievement motives model composed of attitudes towards competition (ST181), motivation to master tasks (ST182), and fear of failure (ST183) scales from the student questionnaire are used in 4-point Likert-type scales such as "strongly disagree (1), disagree (2), agree (3), and strongly agree (4)". While attitudes towards competition and fear of failure scales consist of three items, motivation to master tasks consists of four items (OECD, 2019b). The items of the achievement motives model constructed with these scales, as mentioned in PISA 2018 reports (OECD, 2019a) are shown in Table 2.

Table 2.
The Achievement Motives Model Items

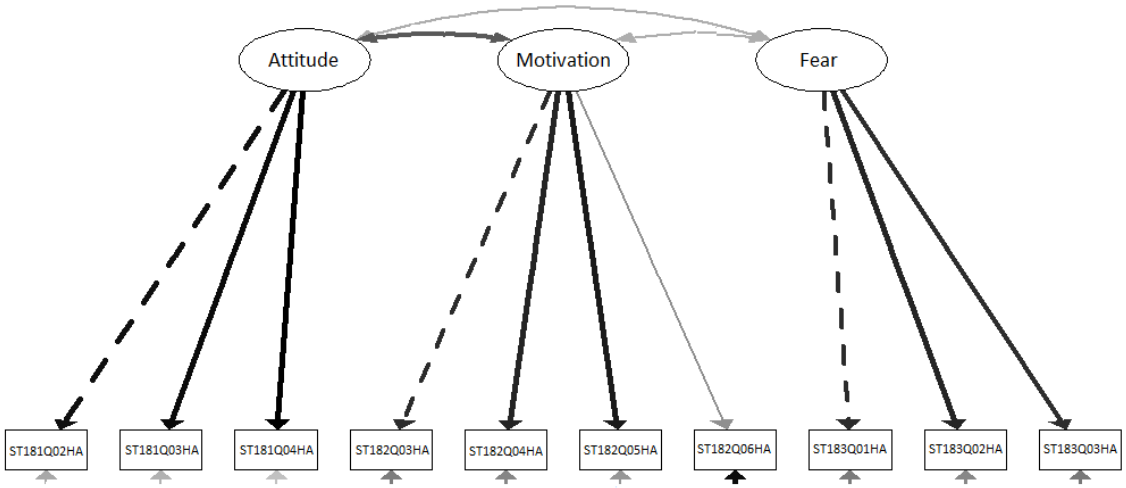
Code of Items	Items
ST181	Attitudes towards competition items
ST181Q02HA	I enjoy working in situations involving competition with others.
ST181Q03HA	It is important for me to perform better than other people on a task.
ST181Q04HA	I try harder when I'm in competition with other people.
ST182	Motivation to master tasks items
ST182Q03HA	I find satisfaction in working as hard as I can.
ST182Q04HA	Once I start a task, I persist until it is finished.
ST182Q05HA	Part of the enjoyment I get from doing things is when I improve on my past performance.
ST182Q06HA	If I am not good at something, I would rather keep struggling to master it than move on to something I may be good at.
ST183	Fear of failure items
ST183Q01HA	When I am failing, I worry about what others think of me.
ST183Q02HA	When I am failing, I am afraid that I might not have enough talent.
ST183Q03HA	When I am failing, this makes me doubt my plans for the future.

Data Analysis

The data were primarily organized and examined to see whether they met the assumptions of the structural equation modeling analysis. The arrangement of the data and the control of the assumptions were made with IBM SPSS Statistics (Version 26). Missing data, outlier values, sample size, multicollinearity, and linearity were examined (Kline, 2015; Tabachnick & Fidell, 2013). Firstly, lower-secondary school data is excluded because of the very limited number of observations (n=22). Remaining cases with missing data were also considered inconsequential because the missing data rate is less than 2% and the missing data is missing completely at random according to the MCAR test (p>.05). Therefore, the listwise method was used (Acuna & Rodriguez, 2004; Kline, 2015, Nakagawa, 2015; Schafer, 1999). Multivariate outliers were computed from the Mahalanobis distance and 129 values were found to show multivariate outliers (p<.001). By excluding individuals with these values from the dataset, analysis was continued with 6442 individuals. It was seen that the dataset obtained from 6442 individuals met the sample size, multicollinearity (examining the variance inflation factor, condition index and tolerance values), and linearity (using scatter plot) and was suitable for SEM analysis. After checking the assumptions, the data were analyzed and the measurement model was established. In this research, measurement invariance of the achievement motives model was examined by MGCFA. The three-factor model analyzed in this study is shown in Figure 1.

Figure 1.

The Achievement Motives Measurement Model



The achievement motives model in which variables were in the specified dimensions was established and it was tested with confirmatory factor analysis (CFA) using SEM to analyze the compatibility of this model with the dataset. The structural equation model was applied with the lavaan package (Rosseel et al., 2022) in R software package (Version 4.0.2). The fit between the model and the data was examined with the goodness of fit statistics. Even though there were several parameter estimation methods for ordinal variables used in CFA/SEM analysis, in this research, the Unweighted Least Squares (ULS) method was used for estimations. The reasons to choose the ULS method include: it is one of the most common methods used for ordinal variables and gives more accurate parameter estimations than diagonally weighted least squares (DWLS) and Maximum likelihood (ML) methods (Forero et al., 2009; Koğar & Yılmaz Koğar, 2015; Yang-Wallentin et al., 2010).

According to the results of CFA, MGCFA was used to determine whether the variables show measurement invariance across gender and school type. Although there are various methods (e.g. alignment method, Bayesian structural equation models) in the examination of measurement invariance, the reason for choosing MGCFA is that MGCFA examines the equivalence of covariance structures, works with latent variables instead of observed variables, and latent means analysis is more sensitive than traditional statistical methods to detect between-group differences (Sehee et al., 2003; Vandenberg & Lance, 2000). Measurement invariance was examined by MGCFA including a sequence of tests of four nested hierarchical models or hypotheses, which are: configural invariance, metric (weak) invariance, scalar (strong) invariance, and strict (residual) invariance (Meredith, 1993;

Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The fit indices of the hierarchically obtained models were examined. While evaluating the fit between the model and the data, the values of chi-square (χ^2), the root mean squared error of approximation (RMSEA; Steiger, 1989), the standardized root mean square residual (SRMR; Bentler, 1995), the comparative fit index (CFI; Bentler, 1990), Tucker–Lewis Index (TLI; Tucker & Lewis, 1973), and the Relative Centrality Index (RNI; McDonald & Marsh, 1990) were taken into consideration.

The value χ^2 is a function of the sample size and tends to reject the null hypothesis when the sample size is large. In other words, the χ^2 test may reject insignificant model-data differences and it is not sufficient by itself (Wu et al., 2007). The deviations of the variables from the normal distribution can inflate goodness-of-fit test statistics (Finney & DiStefano, 2006; Kaplan, 2000). It is thought that it would not be sufficient by itself because the χ^2 is sensitive to sample size and model complexity. For these reasons, Vandenberg and Lance (2000) recommend four indices (RMSEA, SRMR, TLI, and RNI) for overall model fit. RMSEA, SRMR, TLI and RNI are sensitive to misspecified models. SRMR is particularly sensitive to factor covariance misspecification, while others are sensitive to factor loading misspecification. In addition, TLI and RNI are independent of sample size. The reference values for fit indices are stated as follows: .05<RMSEA≤.08 is an acceptable fit, RMSEA≤ .05 is a good fit; .05<SRMR≤.08 is an acceptable fit, SRMR≤.05 is a good fit; .90≤CFI<.95 is an acceptable fit, CFI≥.95 is a good fit; .90≤TLI<.95 is an acceptable fit, TLI≥ .95 is a good fit; .90≤RNI<.95 is an acceptable fit, RNI≥ .95 is a good fit (Hooper et al., 2008; Hu & Bentler, 1999; Tabachnick & Fidell, 2013; Vandenberg & Lance, 2000).

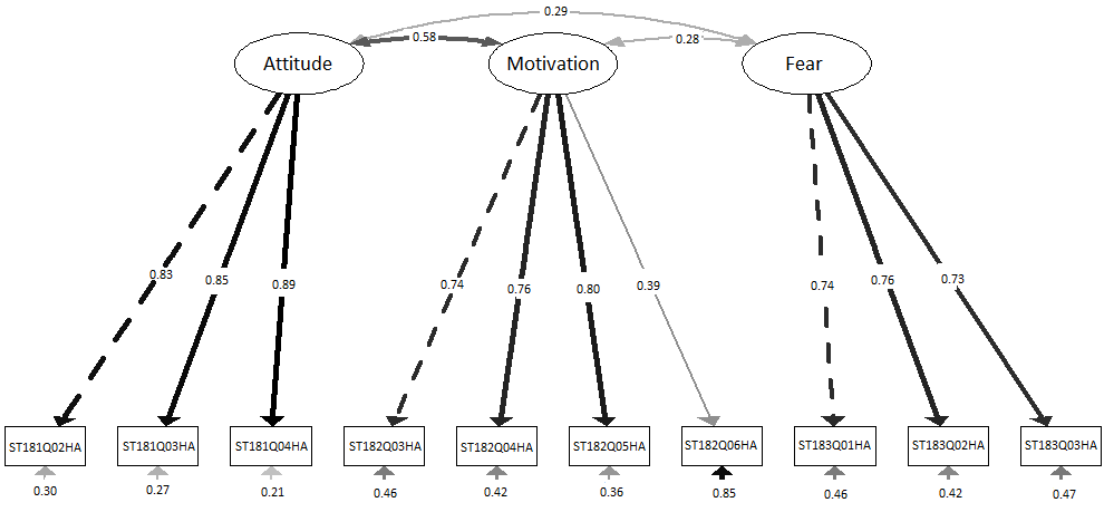
Cheung and Rensvold (2002) state that the likelihood-ratio (LR) test (the chi-square difference test-$\Delta\chi^2$) is generally used to determine model fit differences but $\Delta\chi^2$ test is sensitive to sample size and model complexity as χ^2 test. Yandı et al. (2017) stated that $\Delta\chi^2$ are affected by the degree of freedom and sample size. Dimitrov (2010) indicates that some researchers (e.g., Cheung & Rensvold, 2002; Little, 1997; Vandenberg & Lance, 2000) suggested using changes in other fit statistics to test for measurement invariance because $\Delta\chi^2$ is sensitive to sample size. Şekercioğlu (2018) also agrees that χ^2 is not a practical test for model fit because of statistically sensitive test for large samples and he recommends the use of the most frequently used alternative comparative fit indices like CFI, TLI, and RMSEA instead of χ^2. Cheung and Rensvold (2002) suggest the use of ΔCFI, ΔGamma hat, and ΔMcDonald's Noncentrality Index (ΔMcDonald's NCI) values, which are independent of model parameters and sample size. Furthermore, they indicate the cut-off values as ΔCFI ≤-.01, ΔGamma hat≤-.001, and ΔMcDonald's NCI≤-.02, which means the null hypothesis of invariance should not be rejected. However, Strijbos et al. (2021) state that there is no consensus for the cutoff value for ΔGamma hat and Meade et al. (2006) also state that the value of -.001 may be overly strict because it is affected by small differences in factor loadings. For these reasons, in this study, fit indices (χ^2, RMSEA, CFI, TLI, SRMR, and RNI) of models in addition to the differences of CFI, Gamma hat, and McDonald's NCI values between models are examined to determine measurement invariance. The measurement invariance is tested with the lavaan package (Rosseel et al., 2022) in R software package (Version 4.0.2).

Results

Results on Testing of the Measurement Model

The data were primarily organized and examined to see whether they met the assumptions of SEM analysis as mentioned data analysis (missing data, outlier values, sample size, multicollinearity, and linearity). After checking the assumptions, the three-factor model was established and it was tested with CFA using SEM to analyse the compatibility of this model with the dataset. The model and coefficients obtained according to the results of CFA are given in Figure 2.

Figure 2.
The Achievement Motives Model Path Diagram



The model data fit for the model and subgroups was examined by referring to the indices indicated in Table 3.

Table 3.
Fit Indices of the Achievement Motives Model and Subgroups

Groups	χ^2 (df)	χ^2/df	RMSEA	SRMR	CFI	TLI	RNI
Achievement Motives Model (complete data)	393.129 (32)	**12.285**	.042	.039	.985	.979	.985
Female	161.988 (32)	**5.062**	.036	.037	.986	.980	.986
Male	226.425 (32)	**7.076**	.043	.040	.987	.982	.987
Anatolian High School	187.584 (32)	**5.862**	.041	.041	.982	.975	.982
Vocational and Technical Anatolian High School	123.515 (32)	3.860	.038	.038	.991	.987	.991
Anatolian Imam and Preacher High School	35.207 (32)	3.180	.011	.033	.999	.999	.999
Science, Social Sciences, Multi-Programme Anatolian, Anatolian Sport/Anatolian Fine Arts High School	59.596 (32)	3.930	.034	.046	.989	.985	.989

Note. χ^2 = Chi-square; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Square Residual; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RNI = Relative Centrality Index; Rejections of model invariance were highlighted in bold, **p*<.05.

When the goodness of fit statistics of the scale scores are examined, it is seen that the obtained values show good fits, except χ^2 (χ^2/df = 12.285, *p*<.05; RMSEA = .042, SRMR = .039, CFI = .985, TFI = .979, RNI=.985). This situation can be explained by the sample size and model complexity sensitivity of χ^2. As a result of the analysis, it was determined that the model was compatible with the data because the other fit indices were within acceptable limits. Furthermore, it was seen that the fit between the model and the data across groups was provided (RMSEA≤.05, SRMR≤.05, CFI ≥.95, TFI ≥.95, RNI≥.95).

Results on Testing of the Measurement Invariance by Gender

The measurement invariance of the achievement motives model, which includes three scales, was examined by testing four nested hierarchical models, which are configural invariance, metric invariance, scalar invariance, and strict invariance. Multigroup CFA findings for the three-factor structure equality of the achievement motives model are given in Table 4 according to gender.

Table 4.

Fit Indices for Invariance Tests by Gender Groups

Model	χ2 (df)	RMSEA	SRMR	CFI	TLI	RNI
Configural	388.412 (64)	.040	.036	.987	.982	.987
Metric	438.086 (71)	.040	.038	.985	.981	.985
Scalar	500.368 (78)	.041	.041	.983	.980	.983
Strict	521.799 (88)	.039	.042	.982	.982	.982

Model	Δχ2 (Δdf)	ΔCFI	ΔGamma hat	ΔMcDonald's NCI
Configural	-	-	-	-
Metric	**49.674* (7)**	-.002	-.001	-.003
Scalar	**62.282* (7)**	-.002	**-.002**	-.004
Strict	**21.431* (10)**	-.001	.000	.000

Note. χ^2 = Chi-square; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Square Residual; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RNI = Relative Centrality Index; McDonald's NCI = McDonald's Noncentrality Index; Δ… = Change in fit index. Rejections of model invariance are highlighted in bold, **p*<.05.

The measurement invariance of the achievement model of the PISA 2018 student questionnaire, as shown in Table 4, was established according to gender in the Turkey sample. The four nested hierarchical models were examined and it was seen that the values of the fit indices were acceptable, except χ^2. The differences of χ^2 were significant (p<.05), but it was stated that they should not be evaluated alone because χ^2 is sensitive to the sample size and model complexity in confirmatory factor analytic tests of measurement invariance (Meade et al., 2006). In practice, chi-square is not considered to be a very useful fit index by most researchers because it is affected by several factors (Newsom, 2020). LR tests reject the null hypothesis with too much power if the sample size is large, as the case in our study. In other words, LR tests may reject trivial model-data differences and thus lose practical usefulness (Wu et al., 2007). As the sample size increases, the chi-square value increases, leading to the problem that plausible models are rejected due to trivial discrepancies in measurement invariance tests (Khojasteh, 2012; Wang, 2008; Chen, 2007; Brannick, 1995). Since $\Delta\chi^2$ test is sensitive to sample size, and the sample size in our study (6442) is very high, using the differences in other fit statistics is suggested by the researchers (Cheung & Rensvold, 2002; Şekercioğlu, 2018; Vandenberg & Lance, 2000) to test for measurement invariance. Thus, the goodness of fit indices and ΔCFI, ΔGamma hat, and ΔMcDonald's NCI values were taken as basis in line with the findings of the MGCFA. First, the configural invariance step was provided considering fit indices (RMSEA = .040, SRMR = .036, CFI = .987, TLI = .982, RNI=.987) because the fit indices had acceptable values, except χ^2 ($\chi^2_{(64)}$= 388.412, p<.05). In other words, individuals in different gender groups use the same conceptual framework when answering the scale items. The metric invariance was tested in the second step, and it was observed that metric invariance was held according to fit indices (except χ^2), and the differences of CFI and McDonald's NCI (RMSEA = .040, SRMR = .038, CFI = .985, TLI = .981, RNI=.985; ΔCFI = -0.002, ΔMcDonald's NCI=-.003). On the other hand, LR tests and ΔGamma hat showed that there was no metric invariance ($\chi^2_{(71)}$= 438.086, p<.05; $\Delta\chi^2_7$=49.674, p<.05; ΔGamma hat = -.0013). It was observed that different tests provided different results according to metric invariance. However, as stated above, since the study included a large sample size, LR tests may not provide reliable results due to their sensitivity to sample size. For the ΔGamma hat test, the exact value of -0.0013 was only slightly out of the acceptable range, and it is noted that there is no consensus for the cutoff value for ΔGamma hat (Strijbos et al., 2021). Meade et al. (2006) also state that the value of -

.001 may be overly strict because it is affected by small differences in factor loadings. As a result of these discussions, even though metric invariance was not held based on ΔGamma hat and LR tests, since it was held according to most of the fit indices, ΔCFI and ΔMcDonald's NCI, it was concluded that the factor loadings of the model were the same for male and female groups as in the factor structures of the model. The next step was to check scalar invariance after the configural and metric invariance were found to be satisfied. When the scalar invariance was examined, similar to metric invariance, it was seen that scalar invariance was held according to fit indices (except χ^2), and the differences of CFI and McDonald's NCI (RMSEA = .041, SRMR = .041, CFI = .983, TLI = .980, RNI=.983; ΔCFI = -0.002, ΔMcDonald's NCI=-.004). On the other hand, LR tests and ΔGamma hat showed that there was no scalar invariance ($\chi^2_{(78)}$= 500.368, p<.05; $\Delta\chi^2_7$=62.282, p<.05; ΔGamma hat = -.0017). Due to similar reasons as stated above for the metric invariance, even though scalar invariance was not held based on ΔGamma hat and LR tests, since it was held according to most of the fit indices, ΔCFI and ΔMcDonald's NCI, it was concluded that the regression constants were the same for male and female groups. In the last step, it was observed that the strict invariance was held according to fit indices (except χ^2), and the differences of CFI, Gamma hat and McDonald's NCI (RMSEA = .039, SRMR = .042, CFI = .982, TLI = .982, RNI=.982; ΔCFI = -0.001, ΔGamma hat = -.000, ΔMcDonald's NCI=-.000). On the other hand, LR tests showed that there was no strict invariance ($\chi^2_{(88)}$= 521.799, p<.05; $\Delta\chi^2_{10}$=21.431, p<.05). Thus, because of the same reasons as stated above for the LR tests, since most of the tests agree to have strict invariance, it was concluded that the residual variances for each item are the same in addition to equal factor loadings, slopes and intercepts across groups. Considering the results of the majority of the tests, the full measurement invariance of the achievement motives model is accepted to be held by gender subgroups. As a result, all comparisons made for gender regarding the model will be meaningful according to these findings.

Results on Testing of the Measurement Invariance by School type

Multigroup CFA findings for the three-factor structure equality of the achievement motives model are given in Table 5 according to school type.

Table 5.

Fit Indices for Invariance Tests by School Type Groups

Model	χ2 (df)	RMSEA	SRMR	CFI	TLI	RNI
Configural	405.902 (128)	.037	.037	.989	.984	.989
Metric	443.441 (149)	.035	.038	.988	.986	.988
Scalar	517.921 (170)	.036	.041	.986	.985	.986
Strict	546.103 (200)	.033	.042	.986	.987	.986

Model	Δχ2 (Δdf)	ΔCFI	ΔGamma hat	ΔMcDonald's NCI
Configural	-	-	-	-
Metric	**37.539* (21)**	-.001	.000	-.002
Scalar	**74.480* (21)**	-.002	**-.002**	-.004
Strict	**28.182* (30)**	.000	.000	.000

Note. χ^2 = Chi-square; RMSEA = Root Mean Squared Error of Approximation; SRMR = Standardized Root Mean Square Residual; CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RNI = Relative Centrality Index; McDonald's NCI = McDonald's Noncentrality Index; Δ… = Change in fit index. Rejections of model invariance are highlighted in bold, *p<.05.

The measurement invariance of the achievement model of the PISA 2018 student questionnaire, as shown in Table 5, was established according to school type in the Turkey sample. As mentioned before, the differences of χ^2 were significant (p<.05), but as stated above they should not be evaluated alone because χ^2 is sensitive to the sample size and model complexity, so the fit indices and ΔCFI, ΔGamma hat, and ΔMcDonald's NCI values were also examined. First, the configural invariance step was provided considering fit indices (RMSEA = .037, SRMR = .037, CFI = .989, TLI = .984, RNI = .989) because the fit indices had acceptable values, except χ^2 ($\chi^2_{(128)}$= 405.902, p<.05). In other words, individuals in different school type groups use the same conceptual framework when answering the

scale items. Then, in the second step, the metric invariance was tested, and it was observed that metric invariance was held according to fit indices (except χ^2) (RMSEA = .035, SRMR = .038, CFI = .988, TLI = .986 RNI = .988), and the differences of CFI, Gamma hat and McDonald's NCI (ΔCFI = -0.001, ΔGamma hat = -.000, ΔMcDonald's NCI=-.002). On the other hand, LR tests showed that there was no metric invariance ($\chi^2_{(149)}$= 443.441, p<.05; $\Delta\chi^2_{21}$=37.539, p<.05). As stated above, since the study included a large sample size, LR tests may not provide reliable results due to their sensitivity to sample size. Even though metric invariance was not held based on LR tests, since it was held according to most of the fit indices, ΔCFI, Δ Gamma hat and ΔMcDonald's NCI, it was concluded that the factor loadings of the model were accepted to be the same for school type groups as in the factor structures of the model. The next step was to check scalar invariance after the configural and metric invariances were found to be satisfied. When the scalar invariance was examined, it was seen that scalar invariance was held according to fit indices (except χ^2), and the differences of CFI and McDonald's NCI (RMSEA = .036, SRMR = .041, CFI = .986, TLI = .985, RNI=.986, Gamma hat = .975; ΔCFI = -0.002, ΔMcDonald's NCI=-.004). On the other hand, LR tests and ΔGamma hat showed that there was no scalar invariance ($\chi^2_{(170)}$= 517.921, p<.05; $\Delta\chi^2_{21}$=74.480, p<.05; ΔGamma hat = -.0017). Due to similar reasons as stated above for the scalar invariance considering gender, even though scalar invariance was not held based on ΔGamma hat and LR tests, since it was held according to most of the fit indices, ΔCFI and ΔMcDonald's NCI, it was concluded that the regression constants were the same for school type groups. In the last step, it was observed that the strict invariance was held according to fit indices (except χ^2), and the differences of CFI, Gamma hat and McDonald's NCI (RMSEA = .033, SRMR = .042, CFI = .986, TLI = .987, RNI=.986; ΔCFI = -.000, ΔGamma hat = -.000, ΔMcDonald's NCI=-.000). On the other hand, LR tests showed that there was no strict invariance ($\chi^2_{(200)}$= 546.103, p<.05; $\Delta\chi^2_{30}$=28.182, p<.05). Thus, because of the same reasons as stated above for the LR tests, since most of the tests agree to have strict invariance, it was concluded that the residual variances for each item are the same in addition to equal factor loadings, slopes, and intercepts across groups. Considering the results of the majority of the tests, the full measurement invariance of the achievement motives model is accepted to be held by school type subgroups. As a result, the full measurement invariance of the achievement motives model held by school type subgroups. All comparisons made for school type regarding the model will be meaningful according to these findings.

Discussion

The importance of the individual's affective characteristics in acquiring behaviors and skills in the cognitive domain is known. Affective characteristics also affect school success. Given the role that affective learning outcomes play in shaping students' future behavior, educators should pay attention to students' affective characteristics. Lessons should be developed by taking into account the three learning domains of education, namely cognitive, psychomotor, and affective, and these three domains should be included in the education process. The level of acquisition of these knowledge, skills, and affective characteristics should also be measured and education policies should be planned accordingly. Before making measurements in the affective field, the measurement invariance of the measurement tools to be used must be demonstrated. In this way, it can be determined whether the results obtained are due to the measurement tool or not.

For these reasons, while observing cognitive skills, the individual's affective characteristics should also be taken into account. In international assessment administrations, besides measuring knowledge and skills in cognitive fields, it is also aimed to measure affective characteristics. Some of affective characteristics measured in PISA 2018 are the attitudes towards competition, motivation to master tasks, and fear of failure scales, which are under the achievement motives model.

When it is desired to examine the affective characteristics of individuals or to carry out studies related to these characteristics, first of all, the measurement invariance of the measurement tools that measure these characteristics should be ensured. Measurement invariance is important as it can provide evidence about whether tests/questionnaires measure the same factor in the same way in different groups. In this research, the measurement invariance of the achievement motives model was examined according to gender and school type in the PISA 2018 application in the Turkey sample. The achievement motives model consists of the attitudes towards competition, motivation to master tasks,

and fear of failure scales in the PISA 2018 administration. The three-factor model, established for achievement motives, was tested for the complete datasets, as well as for each gender group and school type group. Confirmatory factor analysis results show that the goodness of fit indices of the measurement model are at acceptable levels except for the lower-secondary school group. Thus, the data of the lower-secondary school group were excluded from the school type dataset. The measurement invariance of the achievement motives model was examined according to gender and school type groups via Multigroup Confirmatory Factor Analysis based on four models. According to MGCFA results, the full measurement invariance of the achievement motives model is accepted to be held by gender and school type subgroups because the values of fit indices and their change are acceptable values, except $\chi 2$. It is noted that both $\Delta\chi 2$ are sensitive to sample size and model complexity (Cheung & Rensvold, 2002; Dimitrov, 2010; Şekercioğlu, 2018; Yandı et al., 2017), and there is no consensus for the cutoff value for ΔGamma hat (Strijbos et al., 2021); the value of -.001 may be overly strict because it is affected by small differences in factor loadings (Meade et al., 2006). Because of these reasons, even though measurement invariance is not held according to LR test results and scalar invariance was not held based on ΔGamma hat, since most of the fit indices, ΔCFI and ΔMcDonald's NCI test results indicate measurement invariance, considering the results of the majority of the tests, the full measurement invariance is accepted to be held according to gender and school type.

Gender differences in achievement motives have been examined in various studies, and there are studies that found differences in achievement motives according to gender (Adsul & Kamble, 2008; Awan et al., 2011; Shekhar & Devi, 2012) as well as studies that do not find any difference (Khan et al., 2011; Yeung et al., 2012; Kaura & Sharma, 2015). In addition, there are some studies examining gender differences in attitudes towards competition, motivation to master tasks, and fear of failure (eber et al.,2015; Eber et al., 2021; Givord, 2020; OECD, 2019b; Severiens & ten Dam, 1998). Before examining the differences by gender, the measurement invariance of achievement motives should be examined. Otherwise, it cannot be determined whether the differences obtained are due to the measurement tool or due to the real differences. The measurement invariance of achievement motives used in the aforementioned studies was examined across genders. Nien and Duda (2008) and Tang and Lu (2013) found that the full measurement invariance held across genders. It can be said that these findings are in parallel with this research. On the other hand, Freund et al. (2011) and Karaman and Smith (2019) found the full measurement invariance is not established across genders.

When the studies examining the achievement motives and their related factors across gender and school type in PISA applications in Turkey are examined, it can be said that these findings are in parallel with the studies by Başusta and Gelbal (2015), Kıbrıslıoğlu (2015), Gülleroğlu (2017), Kıbrıslıoğlu Uysal and Akın Arıkan (2018), and Güngör and Kabasakal (2020) in terms of showing the full measurement invariance of the models according to gender. On the other hand, the studies of Demir (2016), Ardıç and Gelbal (2017), and Uyar and Kaya Uyanık (2019) state that the full measurement invariance is not established according to gender.

Due to the relatively high differences in achievement between school types in Turkey, the results related to the achievement motives model obtained without considering the measurement invariance in the school type may not be valid and reliable (Berberoğlu & Kalender, 2005; Suna et al., 2020). Comparisons by gender or school type will not be meaningful if full measurement invariance is not provided. In this study, the achievement motives model shows full measurement invariance by gender and school type. It can be said that these findings are in parallel with the study by Ardıç and Gelbal (2017), İmrol (2017) in terms of showing the full measurement invariance of the models according to school type, while the study of Uyar and Doğan (2014) does not establish the full measurement invariance according to school type. These results suggest that gender and school type-related measurement invariance merits attention in achievement motives research.

The results of the measurement invariance carried out in this study show that the psychometric qualities of the measurements obtained from the measurement model, which consists of items in the PISA student questionnaire that aim to reveal students' attitudes towards competition, motivation to master tasks, and fear of failure, can be generalized among gender and school type groups. It can be said that the difference between the groups is not due to the measurement tool. The measurements obtained from the achievement motives model items could be generalized among the school groups

and gender, and provide reliable and valid measurements for determining the achievement motives of the students. In this regard, the scores obtained from the achievement model can be used in comparisons according to gender or school type. The results obtained from the competitive attitudes, motivation to master the task, and fear of failure scale can be used reliably and validly to examine the differences between individuals considering gender and school type variables. In addition, the researcher is advised to be cautious when comparing scores in dispositional variables of different groups if there is no evidence about measurement invariance.

This study is limited to the responses given to the achievement motives model in the PISA 2018 student questionnaire towards attitudes towards competition, motivation to master tasks, and fear of failure scale items. In addition, the measurement invariance of the achievement motives model is limited for the group of students at the age of 15 in Turkey. If the achievement motives model is to be used in different age groups, first of all, measurement invariance should be satisfied for that age group, and then the achievement motives model and their scales should be used. The scales or the established models consisting of the scales from PISA should be examined to obtain measurement invariance evidence across groups before using them for the purpose of comparing groups and generalizing the findings. In future studies, researchers can repeat the research using other groups and different models or scales. In addition, measurement invariance studies of the same models can be conducted in different countries. The widely used MGCFA method is used in this study. Other methods can be used and compared to examine measurement invariance because MGCFA has limitations in testing measurement invariance when the number of groups and sample size are large.

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. This study does not require any ethics committee approval as it includes open-access data.

References

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications* (pp. 639-647). Springer, Berlin, Heidelberg.

Adsul, R. K., Kamble, V., & Sangli, K. W. (2008). Achievement motivation as a function of gender, economic background and caste differences in college students. *Journal of the Indian Academy of Applied Psychology*, *34*(2), 323-327.

Ardıç, E., & Gelbal, S. (2017). Cross-group equivalence of interest and motivation items in PISA 2012 Turkey sample. *Eurasian Journal of Educational Research, 17*(68), 221-238. http://dx.doi.org/10.14689/ejer.2017.68.12

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*(6, Pt.1), 359–372. https://doi.org/10.1037/h0043445

Awan, R. U. N., Noureen, G., & Naz, A. (2011). A Study of Relationship between Achievement Motivation, Self Concept and Achievement in English and Mathematics at Secondary Level. *International education studies*, *4*(3), 72-79. http://dx.doi.org/10.5539/ies.v4n3p72

Başusta, N. B & Gelbal, S. (2015). Gruplar arası karşılaştırmalarda ölçme değişmezliğinin test edilmesi: PISA öğrenci anketi örneği. *Hacetepe Üniversitesi Eğitim Fakültesi Dergisi*, *30*(4), 80-90.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.

Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi. *Journal of Educational Sciences & Practices*, *4*(7).

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of organizational behavior, 16*, 201–213.

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural equation modeling: A multidisciplinary journal*, *19*(3), 372-398. http://dx.doi.org/10.1080/10705511.2012.687671

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling: a multidisciplinary journal*, *14*(3), 464-504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Conroy, D. E. (2003). Representational models associated with fear of failure in adolescents and young adults. *Journal of Personality*, *71*(5), 757-783. https://doi.org/10.1111/1467-6494.7105003

De Paola, M., Ponzo, M., & Scoppa, V. (2015). Gender differences in attitudes towards competition: Evidence from the Italian scientific qualification. IZA ddiscussion paper. No: 8859.

Demir, E. (2016). Testing measurement invariance of the students' affective characteristics model across gender sub-groups. *Kuram ve Uygulamada Eğitim Bilimleri, 17*(1), 47-62. https://doi.org/10.12738/estp.2017.1.0223

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121-149. https://doi.org/10.1177/0748175610373459

Ding, Y., Yang Hansen, K., & Klapp, A. (2023). Testing measurement invariance of mathematics self-concept and self-efficacy in PISA using MGCFA and the alignment method. *European Journal of Psychology of Education*, *38*(2), 709-732.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662–680. https://doi.org/10.1037/0021-9010.70.4.662

Eber, N., François, A., & Weill, L. (2021). Gender, age, and attitude toward competition. *Journal of Economic Behavior & Organization*, *192*, 668-690. https://doi.org/ 10.1016/j.jebo.2021.10.022

Education Reform Initative (Eğitimde Reform Girişimi) (2009). *Eğitimde eşitlik: Politika analizi ve öneriler.* ERG Raporları. Retrieved from https://www.egitimreformugirisimi.org/wp-content/uploads/2017/03/Egitimde_Esitlik_Politika_Analizi_ve_Oneriler_1.pdf.

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. Routledge.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock and R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269- 314). Greenwich, CT: Information Age.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A monte carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16, 625 – 641. https://doi.org/ 10.1080/10705510903203573

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Franken, R. E., & Brown, D. J. (1995). Why do people like competition? The motivation for winning, putting forth effort, improving one's performance, performing well, being instrumental and expressing forceful/aggressive behavior, *Personality and Individual Differences, 19*(2), 175-184. https://doi.org/10.1016/0191-8869(95)00035-5

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629-634. https://doi.org/ 10.1016/j.paid.2011.05.033

Givord, P. (2020), "Do boys and girls have similar attitudes towards competition and failure?", *PISA in Focus*, No. 105, OECD Publishing, Paris. https://doi.org/10.1787/a8898906-en.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 Suppl 3), 78-94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Gülleroğlu, H.D. (2017). PISA 2012 matematik uygulamasına katılan türk öğrencilerin duyuşsal özeliklerinin cinsiyete göre ölçme değişmezliğinin incelenmesi. *GEFAD / GUJGEF*, *37*(1), 151-175.

Güngör, M., & Atalay Kabasakal, K. (2020). Investigation of measurement invariance of science motivation and self-efficacy model: PISA 2015 Turkey sample. *International Journal of Assessment Tools in Education*, *7*(2), 207-222. https://doi.org/10.21449/ijate.730481

Hangen E. J., & Elliott A.J. (2016) Achievement motives. In: Zeigler-Hill V., Shackelford T. (eds), *Encyclopedia of personality and individual differences.* (pp. 1-3). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_487-1

Helmreich, R. L., & Spence, J. T. (1978). The work and family orientation questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. *JSAS Catalog of Selected Documents in Psychology*, 8(35), Manuscript No. 1677.

Helmreich, R. L., Spence, J. T., Beane, W. E., Lucker, G. W., & Matthews, K. A. (1980). Making it in academic psychology: Demographic and personality correlates of attainment. *Journal of Personality and Social Psychology*, *39*(5), 896–908. https://doi.org/10.1037/0022-3514.39.5.896

Hooper, D., Coughlan, J. & Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53-60. https://doi.org/10.21427/D7CF7R

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

İmrol, F. (2017). *PISA 2012 Türkiye örnekleminde matematiğe yönelik motivasyon ve öz-inanç yapılarının ölçme değişmezliğinin incelenmesi* [Yüksek lisans tezi]. Ankara Üniversitesi.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409-426. https://doi.org/10.1007/BF02291366

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage

Karaman, M. A., & Smith, R. (2019). Turkish adaptation of achievement motivation measure. *International Journal of Progressive Education*, *15*(5), 185-197. https://doi.org/ 10.29329/ijpe.2019.212.13

Karasar, N. (2019). *Bilimsel araştırma yöntemi: Kavramlar ilkeler teknikler.* Nobel Akademik Yayıncılık.

Kaura, N., & Sharma, R. (2015). The effect of gender on achievement motivation. *Indian Journal of Health & Wellbeing*, *6*(5), 504-507.

Khademi, A. (2020). *An investigation of fit criteria within MG-CFA for examining non-negligible measurement invariance* [Unpublished doctoral dissertation, University of Massachusetts Amherst]. Doctoral Dissertations. https://scholarworks.umass.edu/dissertations_2/2035

Khan, Z., Haider, Z., & Ahmed, N. (2011). Gender difference in achievement motivation of intervarsity level badminton players. *Journal of Physical Education and Sport*, *11*(3), 255.

Khojasteh, J. (2012). *Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in the bifactor model* [Unpublished doctoral dissertation, University of Arkansas]. *Theses and Dissertations*. https://scholarworks.uark.edu/etd/610

Khorramdel, L., Pokropek, A., & van Rijn, P. (2020). Special Topic: establishing comparability and measurement invariance in large-scale assessments, part I. *Psychological Test and Assessment Modeling*, *62*(1), 3-10.

Kıbrıslıoğlu, N. (2015). *PISA 2012 matematik öğrenme modelinin kültürlere ve cinsiyete göre ölçme değişmezliğinin incelenmesi: Türkiye- Çin (Şangay)-Endonezya örneği.* [Yüksek lisans tezi]. Hacettepe Üniversitesi.

Kıbrıslıoğlu Uysal, N, & Akın Arıkan, Ç. (2018). Measurement invariance of science self-efficacy scale in PISA. *International Journal of Assessment Tools in Education*, *5* (2), 325-338. https://doi.org/10.21449/ijate.379508

Kline, R. B. (2015). *Principles and practices of structural equation modeling (4th Ed.).* Taylor & Francis.

Koğar, H., & Koğar, E. Y. (2015). Comparison of different estimation methods for categorical and ordinal data in confirmatory factor analysis. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2). https://doi.org/10.21031/epod.94857

Mascolo, M., & Fischer, K. (1995). Developmental transformations in appraisals for pride, shame and guilt. In J. P. Tangney & K. Fischer (Eds.) S*elf-conscious emotions: The psychology of shame, guilt, embarrassment, and pride* (pp. 64- 113). Guilford.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, *107*(2), 247–255. https://doi.org/10.1037/0033-2909.107.2.247

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2006, August). *The Utility of Alternative Fit Indices in Tests of Measurement Invariance.* Paper presented at the annual Academy of Management conference, Atlanta, GA.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543. https://doi.org/10.1007/BF02294825

MEB (2019). PISA 2018 Türkiye ön raporu. Retrieved January 2024 from, https://pisa.meb.gov.tr/eski%20dosyalar/wp-content/uploads/2020/01/PISA_2018_Turkiye_On_Raporu.pdf

Nakagawa, S. (2015). Missing data: mechanisms, methods and messages. *Ecological statistics: Contemporary theory and application*, 81-105. https://doi.org/10.1093/acprof:oso/9780199672547.003.0005

Newsom, J. T. (2020). Psy 523/623 structural equation modelling, handouts. Some clarifications and recommendations on fit indices. https://web.pdx.edu/~newsomj/semclass/ho_fit.pdf

Nien, C. L., & Duda, J. L. (2008). Antecedents and consequences of approach and avoidance achievement goals: A test of gender invariance. *Psychology of Sport and Exercise*, *9*(3), 352-372. https://doi.org/10.1016/j.psychsport.2007.05.002

OECD (2019a). *PISA 2018 assessment and analytical framework*. OECD Publishing, Paris. https://doi.org/10.1787/b25efab8-en.

OECD (2019b). *PISA 2018 results (Volume II): Where all students can succeed.* OECD Publishing, Paris. https://doi.org/10.1787/b5fd1b8f-en.

Rosseel, Y., et al. (2022). Package 'lavaan'. Documentation available at CRAN: https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. ASA 1988 Proceedings of the Business and Economic Statistics, Section (308-313). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp. 399-419). Thousand Oaks, CA: Sage

Schafer, J. L. (1999). Multiple imputation: a primer. *Stat Methods in Med*, *8*(1), 3–15. https://doi.org/10.1191/096228099671525676.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Sehee H., Malik, M.L., & Lee, M. (2003). Testing configural, metric, scalar, and latent mea invariance across genders in sociotropy and autonomy using a non-western sample. *Educational and Psychological Measurement*, *63*(4), 636-654. https://doi.org/10.1177/0013164403251332

Severiens, S., & ten Dam, G. (1998). A multilevel meta-analysis of gender differences in learning orientations. *British Journal of Educational Psychology*, *68*, 595–608. https://doi.org/10.1111/j.2044-8279.1998.tb01315.x

Shekhar, C., & Devi, R. (2012). Achievement motivation across gender and different academic majors. *Journal of Educational and Developmental Psychology*, *2*(2), 105. https://doi.org/ 10.5539/jedp.v2n2p105

Steenkamp, Jan-Benedict E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78-90. https://doi.org/10.1086/209528

Steiger, J. H. (1989). *EzPATH: Causal modeling.* SYSTAT. https://www.statpower.net/Steiger%20Biblio/EzPath%20Manual.pdf

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), 599–616. https://doi.org/10.1007/s11135-007-9143-x

Strijbos, J. W., Pat-El, R., & Narciss, S. (2021). Structural validity and invariance of the feedback perceptions questionnaire. *Studies in Educational Evaluation*, *68*, 1-13. https://doi.org/10.1016/j.stueduc.2021.100980

Suna, H. E., Tanberkan, H., & Özer, M. (2020). Changes in literacy of students in Turkey by years and school types: Performance of students in PISA applications. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(1), 76-97. https://doi.org/10.21031/epod.702191

Şekercioğlu, G. (2018). Measurement invariance: Concept and implementation. *International Online Journal of Education and Teaching (IOJET)*, *5*(3). 609-634.

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

Tang, H., & Lu, N. (2013). Measure invariance research of the short form of achievement motive scale in boy-girl of middle school students in Shenzhen. *Advances in Psychology*, *3*(6), 321-326. http://dx.doi.org/10.12677/ap.2013.36048

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/BF02291170

Tucker, K. L., Ozer, D. J., Lyubomirsky, S., & Boehm, J. K. (2006). "Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and north Americans": Erratum. *Social Indicators Research*, *78*(2), 341-360. https://doi.org/10.1007/s11205-005-1037-5

Uyar, Ş., & Doğan, N. (2014). PISA 2009 Türkiye örnekleminde öğrenme stratejileri modelinin farklı gruplarda ölçme değişmezliğinin incelenmesi. *Uluslararası Türk Eğitim Bilimleri Dergisi, 2014*(3), 30-43.

Uyar, Ş., & Kaya Uyanık, G. (2019). Fen bilimlerine yönelik öğrenme modelinin ölçme değişmezliğinin incelenmesi: PISA 2015 örneği. *Kastamonu Eğitim Dergisi*, *27*(2), 497-507. https://doi.org/10.24106/kefdergi.2570

van der Sluis, S., Vinkhuyzen, A. A., Boomsma, D. I., & Posthuma, D. (2010). Sex differences in adults' motivation to achieve. *Intelligence*, *38*(4), 433-446. https://doi.org/10.1016/j.intell.2010.04.004

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-69. https://doi.org/10.1177/109442810031002

Wang, C. (2008). *In Search of Diamond Rules* [Unpublished Doctoral dissertation, The Chinese University of Hong Kong]. https://core.ac.uk/download/pdf/48547904.pdf

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation*, *12*(3), 1-26. https://doi.org/10.7275/mhqa-cd89

Yandı, A., Köse, İ. A., & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: PISA 2012 örneği. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 13*(1), 243-253. http://dx.doi.org/10.17860/mersinefd.305952

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling A Multidisciplinary Journal 17*(3), 392 – 423. https://doi.org/10.1080/10705511.2010.489003

Yeung, A. S., Craven, R., & Kaur, G. (2012). Gender differences in achievement motivation: grade and cultural considerations. *Psychology of gender differences*, 59-79.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 35-49

Comparison of Cluster Analysis and Latent Class Analysis for the Detection of Fake Responses on Personality Tests*

İbrahim ŞAHİN ** Seher YALÇIN ***

Abstract

Personality tests reveal the trait being measured, allowing test takers to present themselves differently than they really are. Research suggests that such deception in personality tests can have a negative impact on criterion-related validity. This study compared the effectiveness of cluster analysis and latent class analysis in detecting faking behavior in personality tests. A post-test control group design was used with 543 11th-grade students from eight different high schools in Sanliurfa province during the academic year 2021–2022. Participants in the experimental group were asked to respond in a specific way in order to score higher on the test, believing that their placement in the university depended on the result of the personality test indicating that they had a "positive" profile. Conversely, the control group was asked to present themselves truthfully and give honest answers. In this study, the initial focus was to assess the validity and reliability of the personality test scores. A comparison was then made between the scores of the participants in the experimental and control groups for each sub-dimension of the personality test to determine if there was a significant difference. The findings showed that there was a significant difference in the mean scores between the two groups, with the experimental group having a higher mean score. In addition, the results of cluster analysis and latent class analysis showed that latent class analysis outperformed cluster analysis in detecting fake respondents with a lower error rate.

Keywords: Fake responding, cluster analysis, latent class analysis

Introduction

The decision-making process involves gathering relevant information, comparing it with certain criteria, and reaching a conclusion. Consequently, decision-making can be considered an evaluative process (Turgut & Baykul, 2019). In various stages of education, some decisions need to be made. These decisions may be related to school management, teaching methods, curriculum, selection, placement, classification of individuals, or students' career goals (Thorndike & Thorndike-Christ, 2013). Evaluating data obtained through measurement processes plays a crucial role in determining the effectiveness of educational programs and methods, identifying students' learning deficiencies and achievements, and guiding them toward areas where they can be successful, considering their interests and abilities (Baykul, 2015).

The literature indicates that personality tests are frequently used in research, self-exploration, and clinical decision-making processes. Research purposes for using personality tests include measuring the effectiveness of treatment methods or interventions, helping individuals gain self-awareness under the guidance of a counselor, and making treatment decisions in clinical settings (Thorndike & Thorndike-Christ, 2013). In the field of measurement and evaluation, there has been a focus on examining the applicability of certain assessment tools used in student guidance services as adapted tests in computer-based environments. For example, a self-assessment inventory was employed in one study (Aybek &

*This study was presented at the VIII. International Congress on Measurement and Evaluation in Education and Psychology on September 21-23, 2022. Additionally, this study has been derived from a Master's Thesis conducted under the supervision of Assoc. Prof. Dr. Seher YALÇIN and prepared by İbrahim ŞAHİN.

** Teacher., Republic of Türkiye Ministry of National Education, Şanlıurfa-Türkiye, ibrahimsahinpdr@gmail.com, ORCID ID: 0000-0003-4972-3552
*** Assoc. Prof., Ankara University, Faculty of Education, Ankara- Türkiye, yalcins@ankara.edu.tr, ORCID ID: 0000-0003-0177-6727

To cite this article:
Şahin, İ., Yalçın, S. (2024). Comparison of cluster analysis and latent class analysis for the detection of fake responses on personality tests*, 15*(1), 35-49. https://doi.org/10.21031/epod.1327395

Received: 14.07.2023
Accepted: 4.03.2024

Çıkrıkçı, 2018), while another study used the Skills Confidence Vocational Interest Inventory (Şimşek & Tavşancıl, 2022) to help students recognize their abilities, interests, and values.

Career counseling is one field where personality tests are widely used. Personality traits play a crucial role in various processes, including career choice, career planning, and job satisfaction. When individuals align their career choices with their personality traits, it has a positive impact on their productivity and job satisfaction (Pişkin, 2020). Holland (1973) viewed career choice as a reflection of personality and argued that just as individuals possess personality traits, different occupations also require specific personality traits. Holland categorized these personality traits into six types: realistic, investigative, artistic, social, enterprising, and conventional. According to Holland, individuals tend to gravitate toward professions that allow them to utilize their abilities, attitudes, and values that are consistent with their personality traits. Working in jobs that are compatible with one's personality traits can lead to occupational satisfaction. Törnroos et al. (2019) examined the relationship between personality traits and occupational satisfaction, and their findings are consistent with Holland's perspective. The research indicated that individuals in the same occupation share similar personality traits, and occupational satisfaction increases when there is a match between the average personality traits associated with an occupation and the individual's own personality traits. Moreover, certain personality traits have a greater impact on individuals' occupational choices. The influence of personality traits on job satisfaction and work efficiency emphasizes the importance of accurately measuring personality in career planning and occupational selection. This can be achieved by developing measurement tools that provide valid and reliable results while minimizing the occurrence of fake responses.

Self-report personality inventories have both advantages and disadvantages. On the one hand, individuals themselves are considered to be the best source of accurate information about their own personalities. On the other hand, there are weaknesses associated with this method, such as individuals' lack of sufficient self-knowledge or unwillingness to share certain information about themselves with others. These limitations have led to the need for alternative methods in personality measurement (Cohen & Swerdlik, 2009). In addition, cognitive factors such as inattention, rapid responding, etc., and response styles (such as always tending to give an intermediate response) are also associated with fake responding and inconsistent responding on the scales (Demetriou et al., 2015; Wetzel et al., 2016).

In maximum performance tests, individuals may attempt to give fake responses. In these tests, individuals only have the opportunity to present themselves as less successful than they are. However, in typical response tests, they can present themselves as either better or worse than they truly are (Mehrens & Lehmann, 1991). For instance, when taking an intelligence test, individuals are not expected to perform at a level higher than their current ability, excluding the guessing effect. However, this differs for typical behavioral tests such as personality tests, where individuals with extroverted personalities may intentionally present themselves as introverts. The use of personality tests to measure the suitability of applicants for a job has increased steadily. Meta-analytic studies conducted since the early 1990s have shown that personality tests have an unprecedented level of validity and predictability in personnel selection (Rothstein & Goffin, 2006).

Self-report personality tests operate under the assumption that test-takers will give honest responses, but this assumption may not always be possible. Many respondents may be unwilling to disclose the truth about themselves, even if they are aware of it (Kubinger, 2002). In the context of personnel selection, the use of personality tests is based on two fundamental assumptions. The first assumption is that the instrument effectively measures the intended trait. For instance, if an individual scores high on items measuring honesty in a personality test, it is assumed that he or she is honest in real life. The second assumption is that test scores can predict individuals' future performances. While there is evidence supporting these assumptions, there are valid reasons that remain skeptical about their real-world realization (Adair, 2014). Personality tests often give away the trait they are intended to measure, which allows test takers to present themselves as different from who they are. Research indicates that most job applicants tend to exaggerate their positive traits to increase their chances of being selected, and this deliberate distortion undermines criterion-related validity (Huber, 2017). A meta-analysis by Viswesvaran and Ones (1999) revealed that the personality scores of job applicants were 0.48–0.65 standard deviations higher than those of current employees.

Respondents may also tend to display certain socially acceptable characteristics, even if they do not genuinely possess them. They may prefer to provide responses that ensure social approval rather than reflecting their true views or personality traits (Mehrens & Lehmann, 1991). For example, an individual might respond to items related to freedom of expression in a personality inventory in a way that portrays them as supportive of freedom of expression, even if they are actually intolerant of differing opinions. Some respondents avoid extreme responses and instead opt for moderate responses, making it challenging to gather accurate information from such individuals (Kline, 1999). On the other hand, in personality research, some specifically developed scales were needed to examine the effects of social desirability (Erzen et al., 2021). Additionally, methods such as item response theory models can be used statistically to evaluate the agreement between observed responses and model-predicted responses. These analyzes evaluate how well the model fits the real data and whether this fit is meaningful (Embretson & Reise, 2000). Moreover, statistical methods such as latent profile analysis are also used in the literature to detect classes that react carelessly at the extreme (Maniaci & Rogge, 2014).

The literature highlights the correlation between personality traits and occupational satisfaction (Kang & Malvaso, 2023). When individuals work in jobs that are not aligned with their personality traits, it can have a negative impact on their professional satisfaction and subsequently reduce their productivity. Therefore, the act of faking responses in personality tests should be regarded as more than just an attempt to deceive; it can result in a waste of time and resources. Considering that the process of guiding individuals toward suitable professions begins in secondary education, placing students in university programs that align with their personality traits can lead to a more successful career journey. Consequently, high school students were chosen as the target group for this study. When examining studies conducted in the literature (Huber, 2017; Widhiarso & Himam, 2015; Yankov, 2019), it is seen that the results obtained from the normal process and the directed and encouraged fake responders are compared and that fake responders often have similar response patterns. This study was necessary because of the negative impact of fake responding behavior on the validity and reliability of the scores obtained from the measurement tool, the fact that personality tests are used in important decisions such as hiring individuals, and the importance of detecting intentional errors involved in the measurement process.

In this context, the research aims to address the following sub-objectives:

1. Is there a significant difference in the mean scores of students in experimental and control groups for each sub-dimension of the personality test?
2. To what extent can cluster analysis (CA) detect fake and honest responders in the administered personality test?
3. To what extent can latent class analysis (LCA) detect fake and honest responders in the administered personality test?

Method

A post-test control group design was used in this study. The participants consisted of 11th-grade students from eight different high schools in Sanliurfa during the 2021–2022 academic year. The participants were divided into two groups: the experimental group and the control group. Both groups were administered the Quick Big Five Personality Test (HBBKT). The experimental group was instructed to answer the inventory in a specific way that would present them as the most suitable candidates for admission to a university department, considering that their scores on the inventory would be evaluated for university admission. The control group, on the other hand, was informed that the results of the inventory would be used only for the purposes of a study and were asked to answer the inventory honestly, reflecting their true selves. Assuming that the experimental group gave fake responses and the control group gave honest responses, we examined how accurately the statistical analyses used could classify the respondents. The control group was informed that the results obtained from the inventory would only be used for research purposes and were asked to answer the inventory honestly, reflecting their true selves. No explanation was given to the control group about the experimental design of the study.

Participants

The total initial sample consisted of 705 students, with 363 students in the experimental group and 342 students in the control group. After eliminating data with missing values and extreme outliers during data cleaning, the final sample consisted of 266 students in the experimental group, 277 students in the control group, and 543 students in total. As stated by Dibao-Dina et al. (2014), statistical power is maximum in a sample of equal size. Therefore, the participants in the experimental and control groups were close to each other. Descriptive statistics including the distribution of the study participants by gender are shown in Table 1.

Table 1.

Descriptive Statistics of the Participants

		Experimental Group	Control Group	Total
Gender	Female	130	140	270
	Male	136	137	273
	Total	266	277	543

As seen in Table 1, the distribution of students by gender is relatively similar in both the experimental and control groups. The mean age of the experimental group was 16.1 years with a standard deviation (SD) of 0.36, while the mean age of the control group was slightly higher at 16.38 years with an SD of 0.60. Overall, when considering both groups together, the mean age of the entire study group was 16.24 years with an SD of 0.52.

Data Collection Tool

The Quick Big Five Personality Test (HBBKT) was used in this study. The test, developed by Vermulst and Gerris (2005) and adapted to Turkish by Morsünbül (2014), is based on the Five Factor Theory of Personality. It measures five personality traits: Extraversion, Agreeableness, Emotional Stability, Conscientiousness, and Openness to Experience. The test consists of 30 items, with six items measuring each personality factor. Each item was scored on a seven-point scale. The criterion-related validity of the adapted test was established by examining its relationship with self-concept salience, depression, anxiety, and life satisfaction. The internal consistency reliability of the sub-dimensions of the adapted test ranged from 0.71 to .81, and the test-retest correlation coefficients ranged from 0.80 to .87. In their study, Kutlu and Pamuk (2017) used the adapted test in a Turkish sample of 285 students, reporting Cronbach's alpha values ranging from .69 to .81. Rassart et al (2013) applied the test in a Belgian sample consisting of 366 participants aged between 15 and 20 years, reporting Cronbach's alpha values ranging between .75 and .90. Van der Linden et al. (2010) applied the test to a Dutch sample (mean age 14 years and 10 months) and reported acceptable model-data fit with the following statistics: χ^2=29.24, df=11, NNFI=.92, CFI=.98, RSMEA=.06. They also reported Cronbach's alpha values between .66 and .83.

Data Collection and Analysis

The necessary permissions from the Ministry of National Education and the ethics committee were obtained for the study. The data were collected in November 2021, and the test-retest application was carried out in April 2022. Before implementation, the purpose of the study was explained to school administrators, students, and parents, and informed consent was obtained. We began to prepare the data for analysis by examining missing data. In dealing with nonrandom missing data, it is recommended to delete missing data (Büyüköztürk et al., 2020). For this reason, in this study, data belonging to 60 participants were removed from the dataset as they contained nonrandom missing data.

The second step was to look for outliers. In order to detect univariate outliers, the raw scores in the dataset were converted into standard z-scores. In large samples (n>100), the z-range is accepted as "-4, +4" (Büyüköztürk et al., 2020). In this context, the data of 10 participants with univariate outliers were excluded from the study. In detecting multivariate outliers, the Mahalanobis D^2 statistic is used (Hair et al., 2014). Data from 40 participants with multivariate outliers were deleted. These procedures left data from 266 participants in the experimental group and 277 participants in the control group.

In the third step, the distributions of the data belonging to the experimental and control groups were examined. To do this, the experimental and control groups were considered separately, and the mean, mode, median, standard deviation, kurtosis, and skewness values of each item were checked. Table 2 shows the item statistics for the experimental and control groups.

Table 2.
Item Statistics for Experimental and Control Groups

	Experimental Group					Control Group				
Item	Mode	Median	Mean	Skewness	Kurtosis	Mode	Median	Mean	Skewness	Kurtosis
I 1	4	4	4.25	-0.06	-0.58	6	6	5.84	-0.71	-0.15
I 2	7	6	5.69	-0.48	-0.56	4	4	4.18	0.01	-0.89
I 3	4	4	4.41	-0.21	-0.42	4	4	4.15	0.02	-0.97
I 4	7	5	5.10	-0.55	-0.41	3	4	3.82	0.36	-0.50
I 5	5	5	4.86	-0.53	-0.35	6	6	5.56	-0.52	-0.66
I 6	7	6	6.07	-0.69	-0.34	7	6	5.63	-0.80	0.10
I 7	7	6	5.43	-0.68	-0.29	3	3	3.56	0.16	-0.72
I 8	6	5	5.23	-0.63	-0.26	6	5	5.03	-0.78	.031
I 9	5	5	4.98	-0.62	-0.12	4	4	4.01	0.03	-1.01
I 10	5	5	4.94	-0.72	-0.11	7	6	6.08	-0.73	-0.33
I 11	6	5	5.11	-0.73	0.00	3	4	3.87	0.24	-0.81
I 12	6	6	5.23	-0.77	0.00	6	5	4.82	-0.51	-0.35
I 13	5	5	5.08	-0.73	0.09	4	4	4.04	0.03	-0.99
I 14	5	5	4.95	-0.60	0.16	5	5	5.34	-0.24	-0.31
I 15	6	5	5.17	-0.82	0.16	6	6	5.71	-0.73	0.08
I 16	7	6	6.27	-0.94	0.36	1	3	2.75	0.67	-0.51
I 17	7	6	5.59	-0.96	0.48	4	5	4.64	-0.48	-0.43
I 18	6	6	6.07	-0.93	0.63	5	5	4.83	-0.46	-0.46
I 19	6	6	5.43	-1.06	0.67	7	6	5.52	-0.55	-0.43
I 20	7	6	6.07	-1.00	0.70	7	5	4.82	-0.57	-0.73
I 21	6	6	5.91	-1.06	1.16	3	4	4.04	0.09	-0.96
I 22	7	6	5.77	-1.26	1.31	6	6	5.66	-0.78	0.24
I 23	7	6	6.06	-1.26	1.41	4	4	4.39	-0.24	-0.81
I 24	7	6	5.85	-1.23	1.49	7	6	5.58	-0.73	-0.72
I 25	6	6	5.66	-1.21	1.56	5	5	4.79	-0.48	-0.56
I 26	7	6	6.05	-1.33	1.62	7	5	4.48	-0.14	-1.02
I 27	7	6	6.24	-1.19	1.62	5	5	4.67	-0.56	-0.17
I 28	6	6	5.63	-1.29	1.88	7	6	6.09	-0.94	0.37
I 29	7	6	6.11	-1.53	2.83	1	3	3.23	0.43	-0.66
I 30	7	7	6.25	-2.06	5.77	7	6	5.60	-0.49	-0.64

Analyzing Table 2 separately for the experimental and control groups, it can be seen that the mode, median, and mean values for most items are either equal or closely similar between the two groups. Additionally, the skewness and kurtosis values of all items in the control group fall within the range of ±1, except for the 19th and 28th items in the experimental group, which have skewness and kurtosis values within the range of ±2. Considering the instructions given to the participants in the experimental group, it was expected that this group would have higher scores than the control group. Therefore, these findings are consistent with the objectives of the study. Items 29 and 30 have kurtosis values of 2.83 and 5.77, respectively. It shows that experimental group members gave extreme responses to these items. It may be interpreted as meaning that the students in the experimental group thought that the most important characteristics they should have to be accepted into the university program were the characteristics represented by these items. As a result, the mode, median, and mean values in this group approach 7, indicating a departure from the normal distribution. This suggests that the students followed the given instructions appropriately. In contrast, the data from the control group show a distribution closer to the normal distribution compared with the experimental group, supporting the assumption that the students in the control group gave honest answers in accordance with the instructions given. Considering the data as a whole, it was concluded that the normality assumption was met, allowing the data to be analyzed without any intervention. In addition, the assumption of multivariate normality was examined with Bartlett's Test of Sphericity, and it was concluded that the test result was significant; that is, this assumption was met.

Before the LCA and the CA were carried out, the validity and reliability of the measurement tool were assessed. Table 3 shows the Cronbach's Alpha reliability coefficients for the experimental and control groups.

Table 3.
Cronbach's Alpha and McDonald's Omega Reliability Coefficients for Experimental and Control Groups

	Cronbach's Alpha		McDonald's Omega	
	Experimental Group	Control Group	Experimental Group	Control Group
Agreeableness	.71	.62	.73	.68
Extraversion	.71	.81	.72	.82
Conscientiousness	.79	.80	.80	.81
Emotional Stability	.74	.67	.75	.69
Openness to Experience	.65	.63	.67	.67
Entire Test	.86	.80	.86	.81

When interpreting the calculated Cronbach's Alpha value to assess internal consistency, R. B. Kline (2005) suggests that values of 0.70 and above are considered 'acceptable', .80 and above are considered 'very good', and .90 and above are considered 'excellent'. Additionally, Hair et al. (2014) mentioned that values of 0.60 and above may be acceptable if there is evidence of good construct validity. Nunnally & Bernstein (1994) suggested that McDonald's omega coefficient can be interpreted like Cronbach's alpha, and values above .70 can be considered acceptable. Upon reviewing Table 3, it can be seen that the omega and alpha coefficients of each sub-dimension are close to each other, and all sub-dimensions have reliability coefficients within the acceptable range. Besides, since the reliability coefficients of the control group scores were lower on some subscales, a test-retest method was employed to reinforce the reliability assessment. The first phase of the test-retest was conducted on April 6, 2022, followed by the second phase on April 19, 2022, at Sanliurfa Social Sciences High School, with 39 students participating. The results of the test-retest study are given in Table 4.

Table 4.
Reliability Coefficients for Test-Retest Application

Sub-Dimension	r
Agreeableness	.63
Extraversion	.83
Conscientiousness	.80
Emotional Stability	.74
Openness to Experience	.76

When analyzing Table 4, it can be concluded that there is a strong correlation between the first and second administrations in the sub-dimensions, with the exception of the Agreeableness sub-dimension, where a moderate relationship is observed. Confirmatory factor analysis (CFA) was conducted to verify the original factor structure and assess the measurement tool's construct validity. The software used for CFA was LISREL 8.7, utilizing Maximum Likelihood (ML) Estimation as the estimation method. Prior to the analysis, the dataset was prepared by removing missing data and outliers. The data were then divided into experimental and control groups, and CFA was performed separately for each group. The goodness of fit of the CFA model was assessed based on the $\chi 2/sd$, CFI, RMSEA, and SRMR values. The results of the goodness of fit statistics obtained from the analysis are presented in Table 5.

Table 5.
Confirmatory Factor Analysis Statistics

	χ^2	df	χ^2/df	CFI	RMSEA	SRMR
Experimental Group	800.31	395	2.02*	.93	.062	0.06
Control Group	967.19	395	2.45*	.86	.072	0.07

*p<.001

The first statistic used to assess the model-data fit is the chi-square test. If the chi-square test is not significant, it suggests a good model-data fit. However, this test tends to become significant as the sample size increases (Hair et al., 2014). Therefore, the ratio of the chi-square value to the degrees of freedom, denoted as $\chi 2/df$, can be used as an indicator of model-data fit. A $\chi 2/df$ ratio of 3 or lower indicates a good fit, while a value between 3 and 5 indicates an adequate fit (Sümer, 2000). Examining Table 5, it can be seen that the chi-square tests for both experimental and control groups are significant, but their respective $\chi 2/df$ values are less than 3. This finding indicates a good model-data fit. Another measure used to assess the fit is the Comparative Fit Index (CFI), which ranges from 0 to 1. A CFI value close to 1 indicates a good fit. CFI values of 0.90 or higher are considered acceptable for model-data fit (Westland, 2019). The CFI coefficient of the experimental group exceeded the acceptable level, whereas the CFI value of the control group was close to the acceptable level. Furthermore, a root means square error of approximation (RMSEA) value of 0.05 or lower indicates a good model-data fit (Schumacker & Lomax, 2004). Browne and Cudeck (1993, as cited in Keith, 2015) suggested that RMSEA values of 0.08 or lower are acceptable, whereas values of 0.10 or higher indicate poor model-data fit. In this study, the RMSEA values for both experimental and control groups are within an acceptable range. The SRMR value is interpreted in the same way as RMSEA; therefore, according to SRMR, it can be stated that the model-data fit of both groups is at an acceptable level.

For the first sub-objective of the study, an independent samples t-test was conducted to examine whether there was a significant difference between the participants' scores in the experimental and control groups for each sub-dimension of the personality test. The means of both groups were compared, and the significance of the mean differences was assessed. Additionally, the eta-square effect size was calculated for the significant findings.

For the second and third sub-objectives of the study, the effectiveness of CA and LCA in identifying fake respondents was assessed. Clusters and latent classes obtained from each analysis were named based on the available data, and then the accuracy rates of the analyses were calculated. The correct classification rate is determined by dividing the number of subjects classified as true negative and true positive by the total number of subjects, multiplied by 100 (Hair et al., 2014). In this study, classification accuracies were calculated by dividing the total number of correctly classified participants by the total number of participants.

CA and LCA

CA is a method used to categorize objects based on predetermined criteria, with the goal of identifying the highest similarity within objects and the greatest differentiation between categories. These objects can be respondents to a test, products, or other items under investigation (Hair et al., 2014). In this study, the clustering analysis used the two-step method, which was determined to be suitable for the dataset using SPSS software. The two-step method is designed for large datasets with a predetermined number of clusters and combines hierarchical and nonhierarchical CA techniques (Everitt et al., 2011).

LCA, on the other hand, is a statistical approach that aims to classify individuals into homogeneous subgroups based on their observable response patterns to a series of measurement tools (Geiser, 2013). These latent classes represent unobservable subgroups, where individuals within each subgroup share certain characteristics but differ significantly from individuals in other subgroups (Vermunt & Magidson, 2005). Traditional LCA is similar to CA in that it seeks to identify homogeneous subgroups within a heterogeneous population, often referred to as latent class CA (Vermunt & Magidson, 2002). The data were analyzed using SPSS and Latent Gold software packages.

Results

Results of the Comparison of Scores Achieved by Participants in the Experimental and Control Groups on the Sub-Dimensions of the Measurement Instrument

An independent samples t-test was carried out to assess whether there was a significant difference in the scores obtained by the participants in the experimental and control groups on the sub-dimensions of the measurement tool. The findings of the independent samples t-test are given in Table 6.

Table 6.
Independent Samples T-Test Findings

	Experimental Group (N=266)		Control Group (N=277)		t	Effect Size (η^2)
	Mean	Sd	Mean	Sd		
Agreeableness	36.04	3.97	33.92	4.44	5.85**	.060
Extraversion	29.12	5.93	25.24	7.69	6.58**	.074
Conscientiousness	33.39	5.81	28.10	7.05	9.51**	.143
Emotional Stability	31.47	5.94	23.16	6.27	15.82**	.316
Openness to Experience	35.44	4.17	32.32	4.72	8.15**	.109

**p<.001

When examining Table 6, it is clear that the independent samples t-test conducted for each sub-dimension shows statistically significant results. There was a significant difference in favor of the experimental group across all sub-dimensions. In other words, participants in the experimental group scored higher than the control group on all sub-dimensions. Upon analyzing the effect size values, it can be inferred that the differences in mean scores resulting from group membership are moderate in the sub-dimensions of Agreeableness, Extraversion, and Openness to Experience, whereas they are high in the sub-dimensions of Conscientiousness and Emotional Stability. This indicates that participants in the experimental group portrayed themselves as individuals with more positive traits, aligning with the study objectives. In other words, it shows that the students fulfilled what they were told in the experimental design and that the experimental procedure was effective.

Findings Regarding CA

Within the scope of the second sub-objective of the study, the participants were divided into two groups using cluster analysis. Participants' responses to the test items were used as input for grouping. Since the study consisted of experimental and control groups, the analysis was limited to two groups.

The clusters formed after the analysis were initially labeled as K1 and K2. Separate examinations were made for each sub-dimension, and the groups were named. In this study, the actual group membership of each individual in the clusters is known by the researchers. Therefore, these groups can be named by considering which of the experimental or control groups the majority of individuals in the clusters formed by the analysis are from. It can be said that the new group, consisting mostly of individuals from the experimental group, represents the experimental group, and the other group represents the control group. However, in real life, it remains unclear to which of the groups (fake or honest respondents) the participants belong. Thus, we tried to identify which of the clusters formed by the analysis represents the experimental group and which represents the control group by using information other than the actual group memberships of the individuals. This was done by initially analyzing the size of the clusters. The number of participants in the clusters formed by the analysis and the number of participants in the actual experimental and control groups are summarized in Table 7.

Table 7.
Cluster Sizes Generated by CA

		CA Results		
		K1	K2	Total
Agreeableness	Experimental Group	48	218	266
	Control Group	86	191	277
	Total	134	409	543
Extraversion	Experimental Group	105	161	266
	Control Group	186	91	277
	Total	291	252	543
Conscientiousness	Experimental Group	187	79	266
	Control Group	96	181	277
	Total	283	260	543
Emotional Stability	Experimental Group	191	75	266
	Control Group	50	227	277
	Total	241	302	543
Openness to Experience	Experimental Group	189	77	266
	Control Group	124	153	277
	Total	313	230	543

As shown in Table 7, the sub-dimension of "Conscientiousness" demonstrates the highest similarity between the sizes of the clusters formed during the analysis and the actual group sizes, whereas the sub-dimension of "Agreeableness" exhibits the greatest differentiation. It can be inferred that the Agreeableness sub-dimension has the lowest classification accuracy, even without cluster labeling. Two possible scenarios can arise from this observation. Assuming K1 as the experimental group and K2 as the control group for the Agreeableness sub-dimension, the analysis indicates a higher type two error rate, and vice versa, a higher type one error rate.

Upon examining the dataset, the clusters generated by the analysis for each sub-dimension and the matching rates between the actual experimental and control groups were analyzed. Consequently, it was determined that K1 corresponds to the experimental group and K2 corresponds to the control group for the "Agreeableness" and "Extraversion" sub-dimensions, while the opposite was true for the remaining sub-dimensions. After naming the clusters, the goodness of fit was assessed for each sub-dimension using the chi-square test, and the accurate classification rate was calculated. The reconstructed distribution table, along with the classification accuracy rate and chi-square test findings for each sub-dimension, are given in Table 8.

Table 8.

Classification Accuracy Table regarding CA

		CA Results			Classification Accuracy (%)	Chi-sq Test	
		Exp.	Control	Total		χ^2	df
Agreeableness	Exp.	218	48	266	55.99	12.34**	1
	Control	191	86	277			
	Total	409	134	543			
Extraversion	Exp.	161	105	266	63.90	41.78**	1
	Control	91	186	277			
	Total	252	291	543			
Conscientiousness	Exp.	187	79	266	67.77	69.08**	1
	Control	96	181	277			
	Total	283	260	543			
Emotional Stability	Exp.	191	75	266	76.97	158.84**	1
	Control	50	227	277			
	Total	241	302	543			
Openness to Experience	Exp.	189	77	266	62.98	38.40**	1
	Control	124	153	277			
	Total	313	230	543			

**p<.001

Table 8 shows that the Emotional Stability sub-dimension achieved the highest accurate classification rate in the cluster analysis, with a rate of 76.9%. On the other hand, the Agreeableness sub-dimension had the lowest accurate rate of classification at 55.9%.

The variation in classification accuracy across sub-dimensions can be attributed to several factors. This disparity may stem from the underlying mathematical principles of the analysis itself and potential inconsistencies in participants' adherence to the provided instructions. Even if some participants provided appropriate responses, they might have been assigned to an incorrect cluster. For instance, an individual who genuinely possessed more positive traits and was instructed to respond honestly could have been misclassified as a fake respondent.

Findings Regarding LCA

Within the scope of the third sub-objective of the study, LCA was used to categorize the participants into fake and honest respondent groups based on their responses to the test. A similar approach was adopted as in CA. Initially, the classes generated by the analysis were labeled as S1 and S2. Subsequently, the data were analyzed to determine which class represented the experimental group and which represented the control group. The accuracy of this determination was then confirmed by comparison with the existing dataset. The sizes of the classes formed by the LCA for each sub-dimension

are presented in Table 9, which provides a comparison with the existing experimental and control groups.

Table 9.
Class Sizes Generated by LCA

		LCA Results		
		S1	S2	Total
Agreeableness	Exp.	266	0	266
	Control	63	214	277
	Total	329	214	543
Extraversion	Exp.	193	73	266
	Control	120	157	277
	Total	313	230	543
Conscientiousness	Exp.	86	180	266
	Control	184	93	277
	Total	270	273	543
Emotional Stability	Exp.	209	57	266
	Control	73	204	277
	Total	282	261	543
Openness to Experience	Exp.	104	162	266
	Control	192	85	277
	Total	296	247	543

Upon analyzing Table 9, it becomes clear that the results obtained from LCA closely align with the actual group sizes in the Consciousness and Emotional Stability sub-dimensions. Specifically, in the Conscientiousness sub-dimension, all participants from the experimental group were assigned to the S1 class. This observation without explicitly labeling the latent classes may indicate a high level of accurate classification or possibly suggest the opposite scenario. To gain further insights, the dataset was examined, latent classes were labeled, and their correspondence with the experimental and control groups was comparatively tabulated. Classification accuracy rates were calculated for each sub-dimension, and a chi-square test was conducted. These findings are presented in Table 10.

Table 10.
Classification Accuracy Table regarding LCA

		LCA Results			Classification Accuracy (%)	Chi-sq Test	
		Exp.	Control	Total		χ^2	df
Agreeableness	Exp.	266	0	266	88.40	339.17**	1
	Control	63	214	277			
	Total	329	214	543			
Extraversion	Exp.	193	73	266	64.45	47.50**	1
	Control	120	157	277			
	Total	313	230	543			
Conscientiousness	Exp.	180	86	266	67.03	63.09**	1
	Control	93	184	277			
	Total	273	270	543			
Emotional Stability	Exp.	209	57	266	76.05	148.22**	1
	Control	73	204	277			
	Total	282	261	543			
Openness to Experience	Exp.	162	104	266	65.19	49.96**	1
	Control	85	192	277			
	Total	247	296	543			

**p<.001

Table 10 shows that LCA achieved the highest classification accuracy in the Agreeableness sub-dimension. It accurately classified 88.40% of the participants within this sub-dimension. Furthermore, in the actual application, all participants from the experimental group were correctly classified into the experimental group. The relatively lower rates of the correct classification in other sub-dimensions may be due to inconsistent response patterns among students or the characteristics of the measurement tool employed. Particularly in the Openness to Experience sub-dimension, the presence of inconsistent

responses from students in both the experimental and control latent classes may have led to decreased classification accuracy.

Comparison of CA and LCA

The classification accuracy rates of LCA and CA, as applied for the purposes of this study, are comparatively presented in Table 11.

Table 11.

Classification Accuracy Rates of and LCA

	Classification Accuracy Rate	
	CA (%)	LCA (%)
Agreeableness	55.99	88.40
Extraversion	63.90	64.45
Conscientiousness	67.77	67.03
Emotional Stability	76.97	76.05
Openness to Experience	62.98	65.19

Upon reviewing Table 11, it is evident that LCA achieves a higher accuracy rate for classification in the Agreeableness sub-dimension. In addition, it achieves a nearly equal correct classification rate in the Conscientiousness and Emotional Stability sub-dimensions. CA has the highest accurate classification rate of 76.97% in the Emotional Stability sub-dimension, while its lowest accuracy rate was recorded in the Agreeableness sub-dimension at 55.99%. On the other hand, LCA achieves its highest level of accurate classification rate in the Agreeableness sub-dimension with a rate of 88.4%, while its lowest level is in the Extraversion sub-dimension with a rate of 64.45%. For each analysis, false positive and false negative rates were calculated for each sub-dimension. These rates are shown in Table 12.

Table 12.

False Positive and False Negative Rates for CA and LCA

	CA		LCA	
	False Positive (%)	False Negative (%)	False Positive (%)	False Negative (%)
Agreeableness	68.95	18.05	22.47	0
Extraversion	32.85	39.47	43.32	27.44
Conscientiousness	34.66	29.70	33.57	32.33
Emotional Stability	18.05	28.20	26.35	21.43
Openness to Experience	44.77	28.95	30.69	39.10

Upon examining the false positive and false negative rates of the analyses, it is evident that both analyses exhibit a higher tendency towards false positive. However, in the Extraversion and Emotional Stability sub-dimensions, CA exhibits a higher false negative classification rate. A comparable pattern can be seen in LCA. Here, the false positive classification rate is higher than the false negative classification rate, apart from in the Openness to Experience sub-dimension. These findings suggest that both analyses are more likely to misclassify honest respondents as fake respondents rather than including fake respondents in the honest respondent category.

Discussion and Conclusion

The study initially analyzed the reliability and validity levels of scores derived from the personality test taken by two groups: the experimental group consisting of fake respondents and the control group consisting of honest respondents. Both groups' internal consistency levels were deemed acceptable. Additionally, a test-retest application conducted on the control group revealed moderate stability in the Agreeableness sub-dimension and high stability in the remaining sub-dimensions. The confirmatory factor analysis conducted to assess construct validity yielded goodness-of-fit values that were close to or above the acceptable thresholds. Thus, the construct validity of the scores obtained from the personality test was supported for the study group.

Significant differences were found in the mean scores of the participants between the experimental group and the control group in the personality test, favoring the experimental group. Upon analyzing the effect

size values, it was determined that the level of differences in mean scores resulting from group membership was moderate in the Agreeableness, Extraversion, and Openness to Experience sub-dimensions and high in the Conscientiousness and Emotional Stability sub-dimensions.

When assessing the capacity of CA and LCA to detect fake respondents in the personality test, it was found that LCA exhibited higher classification accuracy in the Agreeableness, Extraversion, and Openness to Experience sub-dimensions, while achieving an equivalent level of accuracy in the Conscientiousness and Emotional Stability sub-dimensions. Consequently, the findings of this study suggest that LCA performed better than CA in detecting fake respondents in personality tests. The divergence in results between the two analyses may be attributed to the mathematical foundations underlying the analyses or the response patterns of the students. This study aligns with the study conducted by Widhiarso and Himam (2015), which examined the detection of fake respondents by CA and LCA. Both studies indicated that CA had a higher frequency of type one errors, while LCA demonstrated higher classification accuracy. Widhiarso and Himam reported classification accuracy rates of 51% to 65% for CA and 55% to 67% for LCA, whereas the current study achieved classification accuracy ranging from 56% to 77% for CA and 65% to 88% for LCA. Thus, the two studies are consistent in terms of which analysis type had higher type one errors and higher classification accuracy. The disparity in classification accuracy levels may be attributed to variations in the study group or the measurement tool used.

Compared to Widhiarso and Himam's (2015) research with a similar objective, this study exhibited higher classification accuracy values in CA. While the prior study achieved its highest classification accuracy in the Openness to Experience sub-dimension, the current study attained the highest accuracy in the Emotional Stability sub-dimension. Both studies consistently indicate that relying solely on CA for the detection of fake respondents is insufficient.

As with CA, LCA yielded higher classification accuracy values compared to the study conducted by Widhiarso and Himam (2015). The prior study reported classification accuracy ranging from 55% to 68%, while the current study achieved values between 65% and 88%. This finding aligns with the outcomes of a study conducted by Magidson and Vermunt (2002) on simulation data with known group memberships, demonstrating that LCA exhibited higher classification accuracy. Given the higher classification accuracy of LCA in the present study, it can be inferred that the findings of both studies are consistent with each other.

Both CA and LCA tend to produce more type one errors than type two errors. In other words, they are more likely to misclassify honest respondents as fake responders. This aspect should be considered during the evaluation process. Additionally, both analyses tend to label individuals with higher mean scores as fake respondents. It is important to keep in mind that individuals with genuinely positive characteristics may be mistakenly labeled as fake respondents by these analyses. Tabachnick and Fidell (2013) argued that the outcomes resulting from type one and type two errors may vary depending on the research objective. In this study, LCA and CA are not considered as methods that can detect fake respondents with complete accuracy but as one of methods that can be used to detect these respondents. Assigning someone who is actually a fake respondent to the honest category by the analysis may lead to this individual not being checked for fake responding. On the other hand, a higher rate of type one error would result in further assessments of individuals who are actually honest respondents, leading to a waste of time and effort. Consequently, it is preferable to have a lower rate of type two errors in this study. Practitioners should consider both situations when making decisions. Furthermore, while it is commonly assumed that individuals' responses to paper-and-pencil measurements are honest and precise, this cannot be conclusively proven by solely relying on such methods. As a solution, it is recommended that researchers employ biometric devices to compare and verify the results of paper and pencil measurements.

In this study, the participants who were instructed to give fake responses were told to think that their admission to a university department would be based on their test scores without specifying which department it was. In future studies, providing a clearer description of the fake personality structure for the group asked to give fake responses may be beneficial. Moreover, this study only examined LCA and CA among the methods used to detect fake responding behavior. Future studies could explore other

analyses and include individuals from different age groups beyond the limited group that participated to this study voluntarily.

Declarations

Conflict of Interest: The authors of the article declare that they have no conflict of interest with any person or organization that may be a party to this study.

Ethical Approval: It is declared that scientific and ethical principles have been followed while carrying out and writing this study and that all the sources used have been properly cited.

References

Adair, C., K. (2014). *Interventions for addressing faking on personality assessments for employee selection: A meta-analysis.* [Doctoral Dissertation, DePaul University], College of Science and Health Theses and Dissertations. https://via.library.depaul.edu/csh_etd/93

Aybek, E. C. & Çıkrıkçı, R. N. (2018). Kendini değerlendirme envanteri'nin bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliği [Applicability of the Self Assessment Inventory as a Computerized Adaptive Test]. *Turkish Psychological Counseling and Guidance Journal, 8* (50), 117-141.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klâsik test teorisi ve uygulaması.* (3. Baskı). Pegem Akademi.

Büyüköztürk, Ş. Şekercioğlu, G. & Çokluk, Ö. (2020). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları.* (6. Baskı). Pegem Akademi.

Cohen, R. J., & Swerdlik, M. (2009). *Psychological testing and assessment* (7th ed.). McGraw-Hill.

Demetriou, C., Uzun Özer, B. & Essa, C. A. (2015). *Self-Report Questionnaires*. The Encyclopedia of Clinical Psychology, First Edition. Edited by Robin L. Cautin and Scott O. Lilienfeld. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118625392.wbecp507

Dibao-Dina, C., Caille, A., Sautenet, B., Chazelle, E., & Giraudeau, B. (2014). Rationale for unequal randomization in clinical trials is rarely reported: a systematic review. *Journal of Clinical Epidemiology, 67*(10), 1070–1075. https://doi.org/10.1016/j.jclinepi.2014.05.015

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates Publishers.

Erzen, E., Yurtçu, M., Ulu Kalın, Ö., & Koçoğlu, E. (2021). Development Of Social Desirability Scale: Validity and Reliability Study. *Electronic Journal of Social Sciences, 20* (78); 879-891. https://doi.org/10.17755/esosder.774947

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). Wiley.

Geiser, C. (2013). *Data analysis with Mplus*. Guilford.

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

Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Prentice-Hall.

Huber, C. R. (2017). *Faking and the validity of personality tests: using new faking-resistant measures to study some old questions.* [Doctoral Dissertation, Minnesota University]. https://hdl.handle.net/11299/185605.

Kang, W., & Malvaso, A. (2023). Associations between Personality Traits and Areas of Job Satisfaction: Pay, Work Itself, Security, and Hours Worked. *Behavioral Sciences, 13*(6), 445. http://dx.doi.org/10.3390/bs13060445

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (2nd ed.). Routledge.

Kline, P. (1999). *The handbook of psychological testing* (2nd ed.). Routledge.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. (2nd ed.). Guilford.

Kubinger, K. D. (2002). On faking personality inventories. *Psychologische Beiträge, 44* (1), 10-16. https://www.proquest.com/docview/212173478

Kutlu, M., & Pamuk, M. (2017). Üniversite öğrencilerinde cep telefonunun problemli kullanımının kişilik bağlamında incelenmesi [Investigation of university students' problematic usage of mobile phone in the context of personality]. *Journal of Human Sciences*, *14*(2), 1263–1272. https://www.j-humansciences.com/ojs/index.php/IJHS/article/view/4073

Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, *20*(1), 36-43.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83.

https://doi.org/10.1016/j.jrp.2013.09.008

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). Wadsworth.

Morsünbül, Ü. (2014). Hızlı büyük beşli kişilik testi Türkçe versiyonu geçerlik ve güvenirlik çalışması [The validity and reliability study of the Turkish version of Quick Big Five Personality Test]. *Düşünen Adam The Journal of Psychiatry and Neurological Sciences, 27*(4), 316-322. https://doi.org/10.5350/dajpn2014270405

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Pişkin, M. (2020). Kariyer gelişim sürecini etkileyen faktörler. B. Yeşilyaprak (Ed.), *Mesleki Rehberlik ve Kariyer Danışmanlığı Kuramdan Uygulamaya* (s. 45-90) içinde. Pegem Akademi.

Rassart, J., Luyckx, K., Goossens, E., Apers, S., Klimstra, T. A., & Moons, P. (2013). *Personality traits, quality of life and perceived health in adolescents with congenital heart disease. Psychology & Health, 28(3), 319 -335.* https://doi.org/10.1080/08870446.2012.729836

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, *16*(2), 155-180. https://doi.org/10.1016/j.hrmr.2006.03.004

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. (2nd ed.). Lawrence Erlbaum Associates.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Türk Psikoloji Yazıları*. *3*(6), 49-74.

Şimşek, A. S., & Tavşancıl, E. (2022). Validity and reliability of Turkish version of skills confidence inventory. *Turkish Psychological Counseling and Guidance Journal*, *12*(64), 89-107. https://doi.org/10.17066/tpdrd.1096008

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

Thorndike, R. M., & Thorndike-Christ, T. M. (2013). *Measurement and evaluation in psychology and education* (8th ed.). Pearson.

Törnroos, M., Jokela, M., & Hakulinen, C. (2019). The relationship between personality and job satisfaction across occupations. *Personality and Individual Differences*, *145*, 82-88. https://doi.org/10.1016/j.paid.2019.03.027

Turgut, M. F., & Baykul, Y. (2019). *Eğitimde ölçme ve değerlendirme* (8. Baskı). Pegem Akademi.

van der Linden, D., Scholte, R. H. J., Cillessen, A. H. N., Nijenhuis, J. t., & Segers, E. (2010). Classroom ratings of likeability and popularity are related to the Big Five and the general factor of personality. *Journal of Research in Personality*, *44*(5), 669–672. https://doi.org/10.1016/j.jrp.2010.08.007

Vermulst, A. A., & Gerris, J. R. (2005). *QBF: Quick big five persoonlijkheidstest handleiding (Quick big five personality test manual).* Leeuwarden: LDC.

Vermunt, J. K., & Magidson, J. (2002). *Latent class cluster analyses*. In J. A. Hagenaars & A. L. McCutcheon (Eds.), Applied latent class analysis (pp. 89-107). Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 user's guide.* Statistical Innovations Inc.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197-210. https://doi.org/10.1177/00131649921969802

Westland, J. C. (2019). *Structural equation models: From paths to networks*. (2nd ed.). Springer.

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). Oxford University Press. https://doi.org/10.1093/med:psych/9780199356942.003.0024

Widhiarso, W., & Himam, F. (2015). Employee recruitment: Identifying response distortion on the personality measure. *Electronic Journal of Business Ethics and Organization Studies, 20*(1), 14-21. http://urn.fi/URN:NBN:fi:jyu-201505061735

Yankov, G. P. (2019). Faking on personality tests: *The relationship between intelligence and personality.* [Doctoral dissertation, Bowling Green State University]. https://scholarworks.bgsu.edu/psychology_diss/206

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 50-64

Do We Really Understand What Formative Assessment Is? Examining the Formative Assessment Definitions Within the Measurement and Evaluation Textbooks*

Seval KULA KARTAL**

Abstract

This study aimed at examining definitions made for the formative assessment within textbooks used for the measurement and evaluation courses of the teacher training programs. It was identified that there are 32 books which are currently accessible and have suitable content for teacher training programs. Based on the measurement and evaluations experts' suggestions and publication dates of the books, the 17 out of 32 textbooks were selected for the study group. It was found out that the 17 textbooks focus on the two themes regarding the formative assessment: how to apply the formative assessment and utilize the outcomes provided by it. The results brought out that the textbooks provide information about various aspects of the formative assessment such as aim, planning, content, application, and feedback process. In addition, the results of the study revealed that textbooks used in teacher training programs for measurement and evaluation courses often contain definitions that include misconceptions and conflicting information compared to the established body of knowledge. This finding indicates that it is required to have textbooks including information on the formative assessment, which is consistent with the recent related literature and cognitive approach. Teachers also need textbooks guiding them towards appropriately applying the formative assessment in the classroom. In addition, instructors are recommended to be aware that most of the textbooks currently utilized in teacher training programs for the measurement and evaluation courses include important misconceptions about formative assessment.

Keywords: formative assessment, misconceptions of formative assessment, educational assessment

Introduction

The concept of student achievement is among the most fundamental concepts within the education field. This concept has been updated in line with the theoretical changes in the fields of psychology and education. Since the 1960s, the concept of student achievement has been redefined aligning with the shift from behaviorism to cognitivism in education. The behaviorist approach accepts that a student is successful if she/he can memorize the information provided to her/him by the teacher and remember the information when needed. However, within the cognitive approach, the concept of student achievement has been updated in a way that students take more active roles in the learning process. The cognitivist theory admits that students are successful if they can apply their knowledge and skills to the problem situations they encounter and self-regulate their learning and motivation processes (Brookhart, 2020; Shepard, 2000).

The change in the definition of student success has required all essential components of the education system to be reconstructed in a way aligns with the updated definition of the concept. The effects of this change have also been observed in the field of measurement and evaluation, which is an important component of the education system (Kula-Kartal, 2022). Before the 1960s, a more conventional approach was dominant in the field of educational measurement. Within that conventional approach, the teaching and learning processes were kept separated from each other, assessing to what extent students can remember the information provided to them became the main aim of classroom assessments and

*Article Turkish Version: https://epodder.org/jmeep/tr/kula-kartal-2024.pdf

** Assoc. Prof., Pamukkale University, Faculty of Education, Denizli-Türkiye, sevalk@pau.edu.tr, ORCID ID: 0000-0002-3018-6972

To cite this article:
Kula Kartal, S. (2024). Do we really understand what formative assessment is? Examining the formative assessment definitions within the measurement and evaluation textbooks. *Journal of Measurement and Evaluation in Education and Psychology,* 15(1), 50-64. https://doi.org/10.21031/epod.1343575

Received: 15.08.2023
Accepted: 5.03.2024

students were compared to each other in terms of their recalling skills (Brookhart, 2020; McMillan, 2020; Shepard & Penuel, 2018). However, since the 2000s, a measurement approach which is more compatible with cognitivism and called formative assessment (FA) or assessment for learning has started to gain wide acceptance within the field (Wiliam, 2011, as cited in Panadero et al., 2018).

The FA is a process during which teachers and students collect information and evidence continuously and systematically with the aim of developing teaching and learning (Popham, 2011; Chappuis, 2009; Chappuis et al., 2013). As one can understand from the definition, assessment is accepted as a dynamic process within this approach. Different from the conventional measurement approach, teaching and learning are integrated processes in the FA. Students rather than teachers are at the center of the whole evaluation process. The FA process focuses on developing and monitoring students' high-order thinking skills instead of evaluating their recalling skills. It aims to improve students' goal-setting, self-monitoring, and decision-making skills. Unlike the conventional measurement approach, the FA puts emphasis on developing students' learning rather than comparing them to each other. To enable that, identifying each student's learning-related weaknesses and strengths and utilizing this knowledge and evidence to provide students with constructive feedback are at the forefront of the FA process.

The FA is a concept which has been in the spotlight of the classroom assessment literature for the last 20 years (Panadero et al., 2018). Nevertheless, some common misconceptions about what the FA is and how it should be applied still exist within the body of literature. Moss and Brookhart (2009) mention the most common three misconceptions about the FA in their book. One of them is to accept the FA as a special test type. The second one is to consider it as an intervention program. Another misconception is to suppose that all measurements providing information to improve teaching or curriculum can be counted as the FA. Similarly, Popham (2011) gave a place to a section focusing on what the FA is not in his book and addressed misconceptions about the concept. For instance, the writer states that instant decisions made by teachers based on their observations of students' behaviors or measuring students' low-level thinking skills frequently during the learning process do not define what the FA is. Supporting those misconceptions about the concept, in a study by Martinez and Martinez (1992; as cited in Black & William, 1998) the FA was conceptualized as testing students frequently. By referring to that study, Black and William (1998) expressed that it can be questioned to define the FA as simply testing students oftentimes. Brookhart and Helena (2003) put forward that the FA is a process integrated with the teaching and learning process and they remark another misconception about the FA, which is defining the FA as applying tests to the students at the end of each teaching unit or section.

The fundamental resources of the classroom assessment literature indicate that there are misconceptions about what the FA is, how it should be applied and how the results obtained from it should be used. The FA requires significant changes in teachers' perceptions about their own and students' roles in classroom assessment applications (Black &William, 1998; Leighton, 2020). However, the textbooks used for teacher training programs are still written under a conventional measurement approach rather than a cognitivist one that may trigger targeted changes in teachers' perceptions. A study by Shepard (2006) provided results supporting this claim. The researcher examined the textbooks utilized for measurement and evaluation courses of the teacher training programs between 1940 and 1990. The headings of chapters within the examined books revealed that the textbooks were written with a very technical and conventional point of view. When Shepard (2020) examined the recent books to explore the current situation, the researcher found out that the textbooks written with behavioristic approach still widely exist, and a limited number of books include chapters focusing on the FA.

In the related body of literature, the concept of FA has been discussed in detail. However, the classroom application of this measurement approach still needs improvement. One reason of this situation is that classroom assessment is still under the effect of the behavioristic measurement approach. This has caused many misconceptions about what the FA is. Shepard (2020) states that it is necessary to take into account this problem observed in the textbooks used for the measurement and evaluation course, which is one of the fundamental courses of the teacher training programs. Teachers' ability to carry out the FA applications appropriately within the classroom depends on teachers' understanding of what the FA is, how to apply it and how to use the results provided by it. Nevertheless, as stated by Black and William (1998), the concept of FA has not been understood properly by most teachers. As a result, classroom

application of the concept still needs development. The misconceptions about the FA were examined in some current international studies (Moss & Brookhart, 2009; Shepard, 2006; 2020; Popham, 2011). However, no study has examined so far how the FA process is defined within the textbooks used for measurement and evaluation courses of teacher training programs in Turkiye. Therefore, the aim of this study is to examine the FA definitions made in the measurement and evaluation textbooks.

Method

Research Model

The current study is a document analysis which aims to figure out how textbooks written in the field of measurement and evaluation define the FA and bring out misconceptions about it. Within document analyses, printed or electronic documents are examined and evaluated based on a systematic process (Bowen, 2009).

Study Group

The study group consisted of 17 textbooks written within the field of measurement and evaluation. To specify the books included in the study group, the keyword "measurement and evaluation in education" was written on the search engine and the results were analyzed. When publishing firms and book-selling websites were reviewed, it was found that there are some books written many years ago in the field of measurement and evaluation and they are not accessible anymore. It was identified that 32 books are currently accessible and have suitable content for teacher training programs after excluding currently inaccessible books. The publication date of the books was used as a criterion in the process of selecting books from the book list and it was seen that 32 books were mostly published between 1981 and 2022. Most of the books were published between 2005 and 2015. The study group was identified after asking two measurement and evaluation experts' opinions regarding to what extent the study group can represent the national measurement and evaluation literature. Accordingly, the two books published between 1980 and 2005, the 12 books published between 2006 and 2015, and the three books published between 2016 and 2021 were included in the study group.

Data Collection and Analysis

In the current study, the document analysis was carried out with the aim of revealing how the FA is defined in the examined textbooks. The document analysis is a process including selecting the data from the documents, evaluating, and synthesizing the selected data. The document analysis process results in sections chosen from the data. The selected sections are organized under main themes and categories by conducting a content analysis process on the selected data (Labuschagne, 2003). The document analysis process of the current study was guided by the following steps suggested by Rapley (2007): a) selecting and generating the document archive based on the research questions, b) skeptically re-reading the document archive for several times, c) coding the documents in a way that will result in as inclusive schemes as possible, d) analyzing regularity and variability observed within the obtained scheme, and e) checking the validity of the results.

By following the steps mentioned above in the current study, the process started with generating a document in which all information given about the FA in the textbooks that was included in the study group. The first reading process of the document was guided by the common misconceptions addressed in the related literature (Moss & Brookhart, 2009; Popham, 2011), which are mentioned in the introduction part of the current study. In the second reading process of the document, different aspects of the FA which were commonly emphasized in the textbooks were entitled and classified under the two themes and seven categories. One of the two themes was the application process of the FA. The five categories revealed under this theme were aim, timing, planning, content of assessments, and classroom applications. The other theme was defined as using the results obtained from the FA process. The focus of the last step was to check the validity of the results. To enable that, the document was critically re-

read by the researcher one more time with the aim of evaluating the appropriateness of defined themes and categories. In addition, the whole document was shared with three measurement and evaluation experts and their opinions were received regarding the appropriateness, comprehensiveness and adequacy of the themes and categories organized by the researcher. The experts were given a form in which there are definitions and statements chosen from the textbooks and exemplifying the related categories entitled by the researcher. The experts were asked to put the statements defining the FA under the category which they find the most related with the relevant definition The Fleiss' Kappa coefficient computed to examine the consistency among experts' opinions was 0,77. The Kappa coefficient indicated that there was a strong agreement among experts regarding to which category given definitions belong. After applying the necessary corrections based on the experts' opinions and recommendations the document analysis process was completed.

Results

The aim of the current study is to reveal the definitions made for the FA in the measurement and evaluation textbooks. After examining the information given for the FA in the textbooks, the aspects that were commonly mentioned within the books were classified under two themes which are given in Table 1.

Table 1.
The Aspects That Are Put Emphasis on In the FA Definitions Within the Measurement and Evaluation Textbooks

The application process of the FA	Aim Timing Planning Content of assessments Classroom applications
Using the results obtained from the FA process	Feedback definition Feedback process

As it can be seen from Table 1, one of the two themes focuses on how the FA is carried out in the classroom. The other theme is about how to utilize the FA results. To examine the definitions covered within the two themes in more detail, firstly, the information given for the aim of the FA in the textbooks was analyzed and the results are given in Table 2.

Table 2.
The Definitions Regarding Aims of The FA Within the Measurement and Evaluation Textbooks

The Aims of FA	**The textbooks**
To reveal which and why learning targets within a learning unit were not acquired by students	A, 1981
To identify students' learning gaps and struggles within each learning unit and providing recommendations to each student to close the gaps	B, 1982
To discover learning gaps, struggles and misconceptions and evaluate the effectiveness of the teaching process	C, 2006
To specify and then remediate students' learning gaps before the teaching process ends	D, 2006
To monitor students' development, reveal their strengths and weaknesses, providing students feedback regarding their development	E, 2006
To identify and remediate the deficiencies occurred during the teaching and learning process	F, 2006
To monitor students' development continuously and identify their learning gaps	G, 2007
To specify and then remediate students' learning gaps	H, 2008
To identify to what extent students have acquired the learning targets and specify their learning difficulties when teaching and learning process is still proceeding, and to evaluate effectiveness of the teaching process	I, 2008
To determine the learning gaps and effectiveness of the teaching process	J, 2010

Table 2.
The Definitions Regarding Aims of The FA Within the Measurement and Evaluation Textbooks (continued)

To specify learning gaps and difficulties	K, 2012
To control whether the teaching program is carried out in line with the targets and plan during the teaching process	L, 2012
To define students' learning gaps, to determine the reasons causing the learning gaps, and take precautions to remediate the gaps	M, 2012
To identify and remediate students' learning gaps	N, 2014
To specify learning gaps and provide students with feedback regarding those gaps	O, 2019
To attain information on which learning targets students have difficulties or gained proficiency, to utilize the information to determine the next steps	P, 2021
To improve students' learning and teaching process, with this aim, to collect information and evidence about students' learning continuously and systematically	R, 2021

When Table 2 is examined, it is seen that one of the aims commonly stated within textbooks is "to identify students' learning gaps." As presented in Table 3 below, within 14 books out of 17 books examined in the current study, the fundamental aim of the FA is accepted as specifying the learning targets on which students have learning gaps. Within the three out of 14 textbooks, to reveal the factors that might cause learning gaps was considered as another aim of the FA. In the eight textbooks, it was stated that the FA also aims at remediating the learning gaps. Lastly, another goal of the FA within the four textbooks was to determine the effectiveness of the teaching process.

Table 3.
The Frequencies of Main Definitions Regarding Aims of the FA Within the Measurement and Evaluation Textbooks

The aims of FA	The textbooks	Frequencies
To identify learning gaps	A, 1981; B, 1982; C, 2006; D, 2006; F, 2006; G, 2007; H, 2008; I, 2008; J, 2010; K, 2012; L, 2012; M, 2012; N, 2014; O, 2014	14
To identify causes of learning gaps	A, 1981; B, 1982; M, 2012	3
To remediate learning gaps	B, 1982; D, 2006; E, 2006; F, 2006; H, 2008; M, 2012; N, 2014; O, 2014	8
To identify effectiveness of teaching process	C, 2006; F, 2006; I, 2008; J, 2010	4
To provide information regarding teaching and learning	P, 2021; R, 2021	2
To develop both teaching and learning	P, 2021; R, 2021	2

According to Table 3, it can be stated that teachers have three aims to accomplish with the FA process: 1) to identify the learning targets on which students have learning gaps, 2) to remediate the learning gaps, and 3) to evaluate their teaching process. The first two of the three aims were commonly stated within most of the examined textbooks. Accordingly, the main aim of the FA is commonly defined as "to identify students' learning gaps and factors causing those learning difficulties" within the textbooks. The two main aims of the FA which are to provide information regarding the teaching and learning and to use that information to develop both teaching and learning were only mentioned within the two most recent books. The results obtained from examining the information about when to apply the FA in the classrooms within the textbooks are given in Table 4.

Table 4.
The Definitions Regarding When to Apply the FA Within the Measurement and Evaluation Textbooks

When to Apply the FA	The definitions	The textbooks	Frequencies
To carry out the FA after completing teaching a unit or topic	After completing teaching units or topics	A, 1981	9
	At the end of units or sections requiring 1-2 weeks or 1-10 hours teaching processes	B, 1982	
	Before completing the instruction or continuing to the next topic	D, 2006	
	At the end of teaching units or after teaching several topics of a unit	F, 2006	
	At the end of a teaching unit and after completing topics.	G, 2007	
	After teaching several learning targets and before starting teaching new targets	J, 2010	
	Teaching and evaluation are integrated. Teachers first teach and then they evaluate what they teach	K, 2012	
	It is applied when teaching and learning activities are continuing. To enable that, it is carried out at the end of a class or unit.	N, 2014	
	It is applied when teaching is continuing. It is more appropriate to apply it after completing teaching a unit or topic.	O, 2019	
To apply FA in a continuous way during the teaching and learning process	When teachers and students are still in the teaching and learning process	P, 2021	2
	The FA is applied in a continuous and systematic way.	R, 2021	
No information	No information regarding when to apply the FA	C, 2006 E, 2006 H, 2008 I, 2008 L, 2012 M, 2012	6

Table 4 reveal that there is no information about when to apply the FA in the six textbooks. In 9 out of the remaining 11 textbooks, it is stated that the FA should be carried out after completing teaching a unit or topic. Within the more recent two textbooks, it is stressed that the FA is required to be carried out in a way that will provide continuous and systematic information when the teaching and learning processes are still going on. Another aspect commonly emphasized in the textbooks is about planning the FA. The information about how to plan the FA in the textbooks is given in Table 5.

Table 5.
The Definitions Regarding How to Plan the FA Within the Measurement and Evaluation Textbooks

Planning the FA	The definitions	The textbooks	Frequencies
Unplanned and informal activities are accepted as the FA.	There are two types of evaluation carried out during the ongoing teaching process. One of them is the instant evaluation carried out by the teachers during the classes. Those instant evaluations help making instant decisions based on informal observations and inferences made for students' learning. Another one is more general. It is planned ahead when to apply it and which targets will be measured by the test.	G, 2007	4
	The FA process includes both the spontaneous evaluation applications and more formal and planned evaluations.	L, 2012	
	The FA can be carried out during the class by utilizing informal evaluation techniques such as questioning, observation. This type of evaluation provides quick and instant information for teachers about the effectivity of teaching.	O, 2019	
	The teaching and assessment process should be planned together. The evaluations which are not planned are informal evaluations. The informal evaluations become meaningful when they provide information regarding the targets acquired by students or the ones on which students need to be developed.	K, 2012	
Teaching and the FA processes are planed separately.	Firstly, all targets covered in the related unit are listed, and then, a test that will measure if each target is acquired by the students is developed.	A, 1981	3
	The targets of the unit aimed to evaluate are identified and listed from the curriculum.	F, 2006	
	After completing teaching a learning unit, if it is needed, the targets of the unit are identified and selected from the curriculum.	J, 2010	
The FA is a previously planned process and inform the daily decisions in the classroom.	The evaluation activities should be previously planned in a way that they can provide information for the daily and weekly decisions of teachers and students.	P, 2021	2
	Teacher should design a planned and systematic FA process and apply it to monitor students' development and provide them on-time feedback.	R, 2021	
No information	No information regarding how to plan the FA.	B, 1982 C, 2006 D, 2006 E, 2006 H, 2008 I, 2008 M, 2012 N, 2014	8

Table 5 reveals that the 8 out of 17 textbooks do not include any information about when or how to plan the FA. When the information given in the remaining nine books was examined, it is understood that the books put emphasis on two points on this issue. One of them is that there are two types of FA which are entitled as formal and informal evaluation. In four textbooks, teachers' questioning, observation and adjusting the teaching based on students' reactions during the class is accepted as a part of the FA.

Within these books, those unplanned and instant activities are entitled as informal evaluations. Another point about planning of the FA within the three textbooks is that teaching and the FA processes are planned separately. According to the information given in those books, teachers first plan the teaching process and apply it. After completing teaching, they plan how and when to assess students. For example, in one of the books, it was stated that "After completing teaching a learning unit, if it is needed, the targets of the unit are identified and selected from the curriculum." This information implies that teachers plan the FA process after completing the relevant teaching unit. The more recent two textbooks emphasized that the FA is planned at the beginning of teaching process so that it can inform daily decisions made by teachers and students in the classroom. The results obtained from examining information about the content of the FA are given in Table 6.

Table 6.
The Definitions Regarding Content of the FA Within the Measurement and Evaluation Textbooks

Content of the FA applications	The definitions	The textbooks	Frequencies
All learning targets covered in a unit should be measured in a unit test in the FA application.	All the learning targets covered by the unit should be measured at the end of the unit.	A, 1981	6
	A unit test should cover all critique learning targets. There should be at least one item measuring each target covered in the unit.	B, 1982	
	It is necessary to evaluate all targets covered in the unit in the FA application.	F, 2006	
	To apply the FA appropriately, all learning targets of the unit should be measured. A follow-up test including many items, requiring students to give short answers, and covering all the targets of the unit should be developed and applied.	G, 2007	
	In a FA application, it is necessary to measure all targets taught within a learning unit.	J, 2010	
	In the FA application, all learning targets covered in a unit are measured.	N, 2014	
The FA applications should focus on monitoring students' higher order thinking skills.	In the FA application, various measurement tools and techniques such as paper-pencil test, projects, performance tasks, and portfolios can be used to obtain information about student's learnings. It is especially recommended to use the FA to monitor skills taking long time to be developed.	R, 2021	1
No information	No information regarding the content of the FA application	C, 2006 D, 2006 E, 2006 H, 2008 I, 2008 K, 2012 L, 2012 M, 2012 O, 2019 P, 2021	10

According to Table 6, the 10 textbooks do not include any information regarding the measurement tools that can be used in the FA process or the content of the tools. Within the remaining seven textbooks, it is reported that all learning targets covered in a unit are required to be measured with at least one item in the FA application. In addition, one textbook emphasizes on using a follow-up test consisting of many items measuring all the targets of the unit and requiring students to give short answers. This information implies that the FA applications are carried out by using a unit test including at least one selective or short-answer item measuring each learning target covered in the unit. Within a more recent textbook, it

is stated that various item types can be used to gain information about students' learning in the FA process. Furthermore, within the same textbook, it is recommended to focus on monitoring higher-order thinking skills taking a long time to develop rather than measuring all the targets in a unit. The information given on how to apply the FA in the textbooks is given in Table 7.

Table 7.
The Definitions Regarding How to Apply the FA in the Classroom Within the Measurement and Evaluation Textbooks

Applying the FA	The definitions	The textbooks	Frequencies
The FA is measurement activities applied frequently during the teaching and learning process.	Teachers can evaluate learning or the effectivity of teaching by repeatedly measuring students' current learnings on the targets.	D, 2006	6
	Students should be evaluated on specific times of the teaching process and there should be short intervals between the evaluations	F, 2006	
	The FA applications should be carried out frequently during the academic year.	J, 2010	
	The FA applied frequently and on time enables developing students' learning. Teacher controls the development of students' learning based on frequent evaluations	K, 2012	
	The learning targets on which students have learning gaps are identified based on the tests applied frequently.	L, 2012	
	The FA is applied more frequently than other types of evaluation.	O, 2019	
The FA is applied by utilizing a unit test.	The results obtained from the follow-up tests applied at the end of a teaching unit are utilized to reveal students' learning gaps and the factors caused those learning difficulties.	A, 1981	11
	The tests used in the FA process are called as formative tests or unit tests.	B, 1982	
	Teacher can evaluate learning or effectivity of teaching by comparing students' learnings in each learning unit during the term.	D, 2006	
	Students are generally evaluated at the end of a teaching unit by using a follow-up test or a unit test.	F, 2006	
	Teacher can evaluate students' learning by utilizing tests called with different names such as learning, unit, formative or pop-up tests.	G, 2007	
	Teachers are required to use measurement tools and techniques like pop-up quizzes, follow-up tests, observation, and interview for the FA	I, 2008	
	After teaching several related learning targets and before starting to teach new ones, a measurement tool measuring students' learnings on each learning target covered in the specific section or unit should be developed.	J, 2010	
	Quizzes and unit tests are the FA.	L, 2012	
	The tests used for FA are called as follow-up or formative tests.	M, 2012	
	The unit tests or quizzes are applied on students for the FA.	N, 2014	
	The first thing to remember when one says the FA is the tests applied at the end of a teaching units.	O, 2019	

Table 7.

The Definitions Regarding How to Apply the FA in the Classroom Within the Measurement and Evaluation Textbooks (continued)

To apply the FA means more beyond than frequently measuring students. The unit tests cannot be defined as FA applications.	The formative power an evaluation applied after completing teaching a unit is weak because learning and teaching process for the relevant unit has already been completed.	P, 2021	2
	To measure students' learning on each unit during the teaching process does not mean that the teacher carried out the FA process. The FA is not to simply measure students' learning frequently.	R, 2021	
No information	No information regarding how to apply the FA	C, 2006 E, 2006 H, 2008	3

Table 7 indicates that the textbooks commonly emphasize the two points about the classroom application of the FA. Within the six textbooks, the FA is defined as measuring students' learning frequently when the teaching and learning process is still going on. Another definition regarding how to apply it is that the FA is associated with utilizing a follow-up test at the end of each unit. 11 out of 17 textbooks include information indicating that the FA means to apply a follow-up test measuring to what extent students have acquired the learning targets covered by the relevant unit. Within two textbooks, it was stated that to apply unit tests at the end of a unit does not mean that the teacher carried out the FA process; on the contrary, the formative power of these kinds of applications will be weak since the teaching and learning processes have already been completed. The information regarding the theoretical definition of feedback within the textbooks is given in Table 8.

Table 8.

The Definitions Regarding the Feedback in the FA Within the Measurement and Evaluation Textbooks

Feedback in the FA	The definitions	The textbooks	Frequencies
Informing students about their learning gaps	To let students learn all learning gaps they have within a unit	A, 1981	9
	Students should be informed about their mistakes and learning gaps. Their mistakes should also be corrected.	B, 1982	
	Teachers can detect what students' learning gaps are on time thanks to the FA.	C, 2006	
	In the FA, not only the learning gaps of individual student but also the learning gaps of whole group are identified.	F, 2006	
	During learning process, students may have learning gaps on some targets. The learning gaps are required to be accomplished. To enable that they must be identified first. The identification of the learning gaps is evaluation.	J, 2010	
	In the FA, not only the learning gaps of individual student but also the learning gaps of whole group are identified.	L, 2012	
	In the FA, not only the learning gaps of individual student but also the learning gaps of whole group are identified.	M, 2012	
	In the FA, not only the learning gaps of individual student but also the learning gaps of whole group are identified.	N, 2014	
	The fundamental aim of the FA is to detect students' learning gaps.	O, 2019	

Table 8.

The Definitions Regarding the Feedback in the FA Within the Measurement and Evaluation Textbooks (continued)

Informing students about both their accomplishments and learning gaps	Test results inform students about what and to what extent they have learned.	E, 2006	4
	The FA applications provide information about students' strengths and weaknesses.	K, 2012	
	The FA provides information for both teachers and students on which targets students are already competent and the ones on which they still need to develop.	P, 2021	
	Not only monitoring but also developing learning is aimed in the FA. To enable that, teachers inform students about both their strengths and weaknesses. In addition, the teacher gives suggestions to students to provide them with guidance towards accomplishing learning gaps.	R, 2021	
No information	No information regarding what feedback is	D, 2006 G, 2007 H, 2008 I, 2008	4

Table 8 reveals that four textbooks do not include any information regarding what feedback is. The feedback is defined with an emphasis on identifying learning gaps within the 9 out of the remaining 13 textbooks. In these textbooks, it is not mentioned to inform students about the targets on which they are already competent. In most of the textbooks, the feedback is defined based on identifying learning targets on which students have difficulties and informing students about their learning gaps. Students' need to be informed about the targets on which they are already competent in addition to the ones on which they need to develop is mentioned within the 4 textbooks. These results indicate that the feedback in the FA process is mostly accepted as informing students about their learning gaps within the examined textbooks. The information included in the textbooks regarding the feedback process is given in Table 9.

Table 9.

The Definitions Regarding the Feedback Process in the FA Within the Measurement and Evaluation Textbooks

The feedback process in the FA	The definitions	The textbooks	Frequencies
The feedback in the FA is a teacher-centered process in which the results provided by the FA are mainly used by teachers.	Teachers should prompt students to participate in supplementary learning activities after they inform students about their learning gaps.	A, 1981	10
	If teachers observe a common learning gap among most of the students, they carry out a supplementary teaching process. If a learning gap is observed among a small group of students, teachers can follow various solutions such as making students do group work or read additional materials.	B, 1982	
	Teachers can take necessary precautions to take account the related factors cause students to have learning gaps when they review their teaching processes based on the results obtained from the FA.	C, 2006	
	Teachers are required to fulfill common learning gaps observed among most of the students.	D, 2006	
	Teachers are expected to fulfill students' learning gaps and consider the effectivity of their teaching processes.	E, 2006	
	Teachers can adjust their teaching plans based on the results obtained from the FA for their later applications. They can make additional activities to fulfill students' learning gaps.	F, 2006	

Table 9.

The Definitions Regarding the Feedback Process in the FA Within the Measurement and Evaluation Textbooks (continued)

	It is teacher's responsibility to identify students' learning gaps. Teachers plan supplementary teaching activities to fulfill the gaps. In addition, they correct their own deficiencies in their teaching plans and applications.	J, 2010	
	Teachers try to fulfill learning gaps by carrying out personal or group-level supplementary teaching activities.	L, 2012	
	The learning gaps identified based on the FA results are fulfilled through additional personal activities or supplementary precautions.	N, 2014	
	In the FA, it is required to identify students' learning gaps and then fulfill those gaps.	O, 2019	
The feedback in the FA is a process in which students can utilize the FA results to self-evaluate and provide feedback to themselves.	Students should be given chances for self-evaluation. Teachers should be in direct and one-to-one communication with the students during the feedback process.	K, 2012	3
	Teachers guide students towards developing their learning by using the results provided by the FA. Students set goals for themselves based on their self-evaluations and teacher's feedback and monitor their own development according to their goals.	P, 2021	
	Student self-evaluate their development based on the results provided by the FA. Teachers inform students about their strengths and weaknesses and suggest students new methods and strategies to develop their weaknesses.	R, 2021	
No information	No information regarding the feedback process	G, 2007 H, 2008 I, 2008 M, 2012	4

Table 9 shows that the four textbooks do not give any information for the feedback process. Within 3 out of the remaining 13 textbooks, it is mentioned that students can utilize the FA results to self-evaluate and provide feedback to themselves. The remaining 10 textbooks have a teacher-centered feedback approach. The feedback process is simply explained in the textbooks as follows: a) the teacher carries out an additional supplementary teaching if a common learning gap is observed among most of the students, b) the teacher prompts students to small group or personal studies and additional readings if a learning gap is observed among a small group of students. Based on the information given for the feedback processes, it is understood that there is a common agreement among textbooks on that the FA results are mainly used by teachers.

Discussion

In the current study, the recent 17 textbooks widely utilized for the measurement evaluation courses of teacher training were examined to reveal the definitions made for the FA. Based on the examination of the books, it was found that the FA is discussed under a behavioristic approach in all books except for the more recent two textbooks. In the textbooks, the FA is defined as follows: the FA is carried out after completing the teaching process of the related unit, a follow-up or unit test is applied to students, this unit test includes items measuring all learning targets covered in the related unit, students are informed about the learning targets on which they have learning gaps, the teacher moves on the next teaching unit if the majority of the class succeeded the test, the teacher carries out additional supplementary teaching activities for the learning targets on which most of the students have learning gaps, and the teacher adapts the teaching process she/he followed for those targets in the future.

As mentioned above, it was revealed that the textbooks put emphasis on how to apply the FA and utilize the results provided by it. In most of the books, the FA is associated with applying unit test on students. This finding is parallel with the related literature. According to Moss and Brookhart (2009), the most common misconception about the FA is to consider it as a test used to reveal what students have learned.

Whereas the FA is a process in which teachers and students collect information with the aim of developing students' learning and adapt their decisions based on that information.

The textbooks commonly state that the FA is carried out after completing the teaching of a unit. This indicates an underlying measurement approach in which teaching and measurement are considered as separate processes. Supportively, Brookhart and Helena (2003) point out that measurement specialists are likely to accept evaluation as something separate from the teaching process. However, the main condition to accept the evaluation as formative is that the evaluation ought to provide information while the learning process is still going on. Therefore, the evaluation should take place in the middle of the teaching and learning process rather than applying it at the end of a learning unit (Shepard, 2000). Chappius et al. (2013), by attracting attention to this misconception, express that the formative power of the evaluation is going to be weak if the evaluation does not meet the two following conditions: 1) the evaluation is carried out on appropriate time that will provide chances for students to take actions, 2) both teachers and students can take actions based on the FA results. The end of a unit is too late to take action because the teaching and learning process of the related unit has already been completed. Supportively, Ferrara et al. (2020) state that when teachers apply unit tests on students, they aim at evaluating to what extent students have learned at the end of the unit rather than targeting to develop teaching and learning.

It is pointed out within the textbooks that there are two types of FA: formal and informal. The instant and spontaneous decisions made by teachers based on their classroom observations are entitled as informal evaluation and accepted as the FA. This acceptance is a misconception about the FA because teachers' instant classroom decisions are not the FA. Supportively, Popham (2011) explains that a teacher can teach a specific concept using a different teaching method if the teacher notices that the majority of students are having difficulty in comprehending the concept based on his or her classroom observations. Those kinds of instant decisions are good since they enable to adapt teaching, but this is not the FA. According to him, the FA is carried out based on a plan instead of instant decisions.

In some textbooks, the FA is defined based on unit tests measuring all the targets covered in the unit and including items requiring students to select or give short answers. It can be stated that there are two misconceptions regarding the content and application of the FA within the books. One of them is to focus on students' recalling skills. Popham (2011), drawing attention to this misconception, considers that it is meaningful to utilize the FA to monitor students' skills taking a long time to be develop such as critical thinking or problem-solving. Another misconception is to consider that the FA includes only traditional item types such as multiple-choice items. Whereas the FA process comprises all measurement methods (performance tasks, portfolio etc.) providing necessary information about students' learning development (Kula-Kartal, 2021).

The sections explaining how to use results provided by the FA within the textbooks mostly focus on what feedback is. The feedback is defined based on informing students about their learning gaps in most of the books. From this point of view, the feedback means checking how much information or concepts can be recalled or comprehended by students. In addition, the feedback simply means to inform students about their correct and wrong answers. Shepard (2000) accepts that the behavioristic approach caused unsophisticated feedback definitions observed in the books. In the FA, the fundamental aim is to develop students' learning. However, to inform students about the score they need to gain to be accepted as competent on a learning target is not a type of feedback that can help them develop their learning. To help students develop their learning, it is necessary that the feedback should answer the questions asked by teachers and students like "What are the learning targets? To what extend have we progressed towards the targets? What should be done for a better progress?" during the teaching and learning processes (Hattie & Timperley, 2007). In formative feedback, teachers compare students' performances with criteria defining expected performance. They identify which criteria are accomplished and which ones are not met by the performance. They suggest new methods to develop weak aspects of students' performances. Thus, the feedback enables students to have a view regarding their own performance and creates an opportunity for development (Moss & Brookhart, 2009).

Within the textbooks, the dominant perspective is that teachers are mostly the ones who utilize the results provided by the FA. The teacher has the role of making decisions to develop teaching or learning

based on the assessment results in most of the books. However, some researchers think that the teacher-centered feedback process is not appropriate for the FA. For example, Brookhart and Helena (2003) express that the following conditions should be met by the feedback to be accepted as formative feedback: Firstly, students should comprehend what the criteria defining the expected performance mean. Secondly, students should monitor their own performance and compare it with the criteria. Thirdly, they should take action to lessen the gaps between their performances and criteria. Student is in the center of this formative process because she/he is the only person that can take the necessary actions to develop learning. To help students evaluate their own performances and provide themselves with internal feedback rather than providing external feedback to them form the foundation of the FA. Therefore, self-evaluation is an important component of the FA (McMillian, 2020; Panadero et al., 2018).

It was aimed to examine definitions made for the FA in the measurement and evaluation textbooks within the current research. It was found that the 17 textbooks examined focus on two themes: how to apply the FA and utilize the results provided by it. The results brought out that the definitions provide information about various aspects of the FA such as aim, planning, content, application, and feedback process. In addition, the results of the current study revealed that in the textbooks used for measurement and evaluation courses of teacher training programs, there are some definitions including misconceptions and conflicting information with the related research as discussed in this section. This finding indicates that it is required to have textbooks including information regarding the FA that is consistent with the recent related literature and cognitive approach. Teachers also need textbooks guiding them towards appropriately applying the FA in the classroom. In addition, textbooks and sources including appropriate information about the FA can contribute to both measurement and evaluation literature and teachers' classroom applications. In addition, instructors are recommended to be aware that most of the textbooks currently utilized for the measurement and evaluation courses of the teacher training programs include important misconceptions about the FA. Therefore, it is important to critically review the sources they used for their courses.

Declarations

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

Ethical Approval: The published textbooks were examined in this study. Therefore, ethical approval is not required.

References

Black, P., & William, D. (1998). Assessment and classroom learning. assessment in education: principles, policy & practice, 5(1), 7-74. https://doi.org/10.1080/0969595980050102

Bowen, G.A. (2009). Document analysis as a qualitative research method. Qualitative Research Journal, 9(2), 27-40. https://doi.org/10.3316/QRJ0902027

Brookhart, S. M (2020). Feedback and measurement. S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 63-78). Taylor & Francis.

Brookhart, S. M. & Helena, M. T. (2003). Developing measurement theory for classroom assessment purposes and uses. Educational Measurement: Issues and Practice, 22(4), 5-12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Chappuis, J. (2009). Seven strategies of assessment for learning. Pearson Education.

Chappuis, J., Stiggins, R., Chappuis, S., & Arter, J. A. (2013). Classroom assessment for student learning. Pearson Education.

Ferrara, S., Maxey-Moore, K., & Brookhart, S. M. (2020). Guidance in the standards for classroom assessment: useful or irrelevant? S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 97-119). Taylor & Francis.

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487

Kula-Kartal, S. (2021). Performans görevlerine ve portfolyoya dayalı durum belirleme. M. D. Şahin (Eds.), Eğitimde ölçme ve değerlendirme. (p. 175-206). Nobel Akademik Yayıncılık.

Kula-Kartal, S. (2022). Classroom assessment: The psychological and theoretical foundations of the FA. International Journal of Assessment Tools in Education, 9(Special Issue), 19-27. https://doi.org/10.21449/ijate.1127958

Labuschagne, A. (2003). Qualitative research-airy fairy or fundamental?. The Qualitative Report, 8(1), 100-103. https://doi.org/10.46743/2160-3715/2003.1901

Leighton, J. P. (2020). Cognitive diagnosis is not enough: the challenge of measuring learning with classroom assessments. S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 170-191). Taylor & Francis.

McMillan, J. H. (2020). Assessment information in context. S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 79-94). Taylor & Francis

Panadero, E., Andrade, H., & Brookhart, S. M. (2018). Fusing self-regulated learning and FA: a roadmap of where we are, how we got here, and where we are going. The Australian Educational Researcher, 45, 13-31. https://doi.org/10.1007/s13384-018-0258-y

Popham, W. J. (2011). Classroom assessment: What teachers need to know. Pearson.

Rapley, T. (2007). Doing conversation, discourse and document analysis. SAGE Publications.

Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14. https://doi.org/10.3102/0013189X029007004

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 623–646). Westport, CT: Greenwood Publishing Group.

Shepard, L. A. (2020). Discussion of part II: Should "measurement" have a role in teacher learning about classroom assessment? S. M. Brookhart & J. H. McMillan (Eds.), Classroom assessment and educational measurement. (p. 192-206). Taylor & Francis.

Shepard, L. A. & Penuel, W. R. (2018). Using learning and motivation theories to coherently link FA, grading practices, and large-scale assessment. Educational Measurement: Issues and Practice, 37(1), 21-34. https://doi.org/10.1111/emip.12189

Wiliam, D. (2011). What is assessment for learning? Studies In Educational Evaluation, 37(1), 3-14. https://doi.org/10.1016/j.stueduc.2011.03.001

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 65-78

Investigating the Effect of Testlets Consisting of Open-Ended and Multiple-Choice Items on Reliability via Generalizability Theory*

Serpil KOCAOĞLU ** Melek Gülşah ŞAHİN ***

Abstract

This study aimed to reveal the effect on reliability of testlets consisting of open-ended and multiple-choice items with similar content. For this purpose, two different mathematics achievement tests, one with multiple-choice items and the other with open-ended items, were administered to 128 8th-grade students. Reliability estimations on the obtained data were conducted in the Edu-G program based on the Generalizability Theory. A decision study was also performed. In the achievement test with testlets consisting of open-ended items, p×i×r (p: person, i: item, r: rater) fully crossed design was used when testlet effect was not considered; p×(i:t)×r (t: testlet) nested design was used when testlet effect was considered. According to the results, the reliability coefficient was estimated higher when the testlet effect was not considered. Similarly, in the achievement test with testlets consisting of multiple-choice items, the p×i crossed design was used when the testlet effect was not considered, and the p×(i:t) nested design was used when the testlet effect was considered. According to the results, the reliability coefficient was estimated higher when the testlet effect was not considered. According to the data obtained within the scope of the study, the reliability coefficient was estimated higher in the test with open-ended items than in the test with multiple-choice items. When the testlet effect was included, the change in the reliability coefficient in the test with open-ended items was higher than the change in the test with multiple-choice items. In the decision studies, it was observed that the increase in the number of items and testlets positively affected reliability, but the increase in testlets contributed to reliability more. In the tests consisting of open-ended items, it was observed that the increase in the number of raters contributed to reliability less than items and testlets.

Keywords: Open-ended items, multiple-choice items, testlet, generalizability theory, reliability

Introduction

Assessments are carried out for different purposes at every stage of the education processes. Recognizing students, identifying and eliminating students' learning deficiencies, organizing learning experiences, determining students' learning levels, organizing the learning environment, etc., can be stated within these purposes. In the assessment process, firstly, the appropriate measurement tool should be selected, and the measurement process should be planned. When measurement tools are examined in general, it can be stated that they have different characteristics. Multiple-choice items, the most effective and useful way of measuring knowledge (Haladyna, 2004), are frequently preferred in national and international examinations. Multiple-choice tests are objective tests in terms of scoring. The data obtained by applying multiple-choice tests to a large number of groups can be evaluated in a short time. Scoring is also easy and takes a short time. It can be prepared at all levels of the cognitive taxonomy (Downing, 2006). However, measuring high-level cognitive skills in assessments conducted with multiple-choice tests is difficult (Ko, 2010). Open-ended items can especially be used in the measurement of high-level skills. Open-ended items have three important advantages over multiple-choice items (Bridgeman, 1992). With open-ended items, there is no possibility of finding the correct

*This study has been produced from Master's Thesis that was conducted under the supervision of the Assoc. Prof. Dr. Melek Gülşah ŞAHİN and prepared by Serpil KOCAOĞLU.

** Teacher, Republic of Türkiye Ministry of National Education, Ankara-Türkiye, zncl10@gmail.com, ORCID ID: 0009-0008-0566-4371

***Assoc. Prof. Dr., Gazi University, Gazi Faculty of Education, Ankara-Türkiye, mgulsahsahin@gazi.edu.tr, ORCID ID: 0000-0001-5139-9777

To cite this article:
Kocaoğlu, S. & Şahin, M.G. (2024). Investigating the effect of testlets consisting of open-ended and multiple-choice items on reliability via generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 15*(1), 65-78. https://doi.org/10.21031/epod.1429423

Received: 1.01.2024
Accepted: 25.03.2024

answer by guessing. In addition, although it is not intended as feedback, individuals may realize that they have made a mistake when they cannot find what they think as the answer in the options of multiple-choice items; however, such feedback is not possible in open-ended items. Thirdly, it is impossible to reach the correct answer in open-ended items by eliminating the options as in multiple-choice items.

In addition, in multiple-choice items, the individual can find the answer with less effort compared to open-ended items (Attali et al, 2016). While students are sometimes expected to give a short and strictly defined answer to the items, sometimes the student can be left free in terms of the quality and length of the answer. When considered in this context, open-ended items are categorized under two headings: restricted and unrestricted response (Berberoğlu, 2006). In restricted response questions, the answers are mostly short since some limitations are imposed on the quality or length of the answer. On the other hand, in free-response questions, since the respondent is given a certain amount of freedom regarding the quality or length of the answer, these questions are long-answer questions (Doğan, 2009). In addition to the multiple-choice and open-ended items mentioned here, item types such as short-answer, fill-in-the-blank, true-false, and matching are frequently used in the literature (Doğan, 2019a; 2019b; Karatoprak Erşen & Gündüz, 2023; Nitko & Brookhart, 2014; Popham, 2014; Russell & Airasian, 2008).

In the selection of the item type, the purpose of the test, the feature to be measured, the group to be measured, the application conditions, etc., should be taken into consideration. Another issue that should also be considered is how the items will be presented. When the use of items in both national and international exams is examined, it is seen that items are presented independently or in testlets. The concept of testlets was first introduced by Wainer and Kiely (1987) to refer to a group of items with a common stimulus. In their study, they used testlets in Computerized Adaptive Testing (CAT) to balance the content and eliminate the importance of context effect and item order. Testlets have been widely used especially in recent years. Item writers prefer to prepare items based on a common material because it saves time and energy (Wainer et al. 2000). Moreover, it has been observed that following consecutive items based on a common root is more successful (Lee et al, 2000). The fact that it makes it possible to measure high-level cognitive skills is effective in making testlets popular. In a testlet, a common material such as a graph, table, figure, or map is used to answer two or more items. Some rules, such as the comprehensibility of the material, the ability to respond to the items correctly based only on the material, and careful determination of the number of items, should be considered in developing testlets (Tekin, 2009). Although there is no definite rule in determining the number of items that should be included in testlets (Kaya Uyanık & Ertuna, 2022), there may be between 2-12 items (Yaman, 2016) depending on the characteristics of the common material. In determining the number of items in testlets, the characteristics of the structure shared by the items and content validity can also be taken into consideration.

The limitation is that the items in testlets have local dependency on each other. Therefore, the testlet effect should be considered in estimating the test reliability in which testlets are included. When the items in the testlets are considered independently, the reliability value can be estimated to be higher than when the testlet effect is taken into account. (Lee et al, 2000; Sireci et al, 1991; Taşdelen Teker, 2014; Wainer & Thissen, 1996).

This study aimed to estimate the reliability of testlets consisting of open-ended items and multiple-choice items prepared in similar content within the framework of generalizability theory in cases where the testlet effect was taken and not considered. In recent years, it is seen that the frequency of testlets has increased both in Türkiye's national exams such as the Academic Personnel and Postgraduate Education Entrance Exam (ALES), Foreign Language Exam (YDS), Higher Education Institutions Foreign Language Exam (YÖKDİL) and in international exams such as Test of English as a Foreign Language (TOEFL) and Program for International Student Assessment (PISA). Although using testlets consisting of multiple-choice items is preferred chiefly due to the scoring advantage, the use of testlets consisting of open-ended items cannot be ignored. In this context, the reliability estimations of testlets with similar content prepared with different item types will shed light on the researchers who would like to work on this subject. The study used the Generalizability (G) theory to determine the reliability estimation. According to Shavelson and Webb (1991), G theory is a statistical theory that gives an idea

about the reliability of behavioral measurements. As an extension of both Classical Test Theory and analysis of variance, G theory is a mathematical model in which multiple sources of error can be addressed. The advantage of G-theory is that different error sources can be estimated simultaneously with a single analysis. In other words, unlike Classical Test Theory, it considers the results of different error sources separately and in a single interaction. G theory also allows us to estimate the reliability of scores for different interpretations. While in Classical Test Theory, only relative assessments are made in which individuals are compared with each other, in G theory, it is also possible to make absolute assessments in which only the performance of individuals is evaluated independently of each other. In other words, while Classical Test Theory provides researchers with information to make only relative decisions, G theory offers sufficient information for both relative and absolute decisions at the same time (Brennan, 2001; Güler et al, 2012; Shavelson & Webb, 1991). Within the scope of the study, reliability estimation was performed using the Generalizability approach, and a decision study was conducted.

In the literature, there are studies in which testlets are handled with Item Response Theory (Sireci et al, 1991; Wainer & Thissen, 1996) and G theory (Lee & Frisbie, 1999; Lee et al, 2000; Taşdelen Teker, 2014; Kaya Uyanık & Gelbal, 2018; Kaya Uyanık & Ertuna, 2022). In addition, while Gessaroli and Folske (2002) addressed testlets with factor analysis, Kaya Uyanık and Gelbal (2018) studied two-facet patterns with the Generalizability approach in Item Response modeling using testlet data generated in simulation in their study and compared the results obtained with the results from G theory. Although it was seen that the studies conducted in the international arena occupied a larger space, it was seen that the studies conducted in the national arena were limited. In addition, there is no study in the literature examining the reliability of testlets consisting of both open-ended items and multiple-choice items with similar content within the framework of G theory. From this point of view, it is thought that this study will also contribute to the literature. In the study, a decision (D) study was conducted on the effect of the number of items and the number of testlets on reliability estimation in testlets consisting of open-ended and multiple-choice items. In addition, a D study was also conducted for the change in the number of raters for the test composed of open-ended items where more than one rater was involved. For this reason, the study is considered to be vital as it will provide a different suggestion to the users in exams where testlets are frequently used. The research problems formed in line with the purpose of the study were determined as follows:

1. In achievement tests with testlets consisting of open-ended items;
    a. What are the variance components and G and Phi coefficients for the p×i×r fully crossed design in which person (p), item (i), and raters (r) are crossed with each other when the testlet effect is not considered?
    b. What are the G and Phi coefficients for the decision studies on increasing or decreasing the number of items and raters in the p×i×r design in which the testlet effect is not taken into account?
    c. What are the variance components and G and Phi coefficients of the p×(i:t)×r design in which items (i) are nested in testlets (t) and individuals (p) and raters (r) are crossed with them?
    d. What are the G and Phi coefficients of the p×(i:t)×r design in which the testlet effect is taken into account in the decision studies for increasing or decreasing the number of testlets, items in the testlets, and the raters?
2. In achievement tests with testlets consisting of multiple-choice items;
    a. What are the variance components and G and Phi coefficients of the p×i design in which persons (p) are crossed with items (i) when the testlet effect is not considered?
    b. What are the G and Phi coefficients in the decision study for increasing or decreasing the number of items in the p×i design where the testlet effect is not considered?
    c. What are the variance components and G and Phi coefficients of the p×(i:t) partial nested design in which items (i) are nested in testlets (t) and persons (p) are crossed with them?
    d. What are the G and Phi coefficients for decision studies with increasing and decreasing the number of testlets and the number of items within a testlet in the p×(i:t) design where the testlet effect is taken into account?

Method

This study aimed to obtain and compare G and Phi coefficients within the framework of Generalizability theory in achievement tests with testlets consisting of multiple-choice and open-ended items in cases where the testlet effect was and was not taken into account. The research is basic research since it aims to obtain new information by testing the existing theory in different situations (Karasar, 1994).

Participants

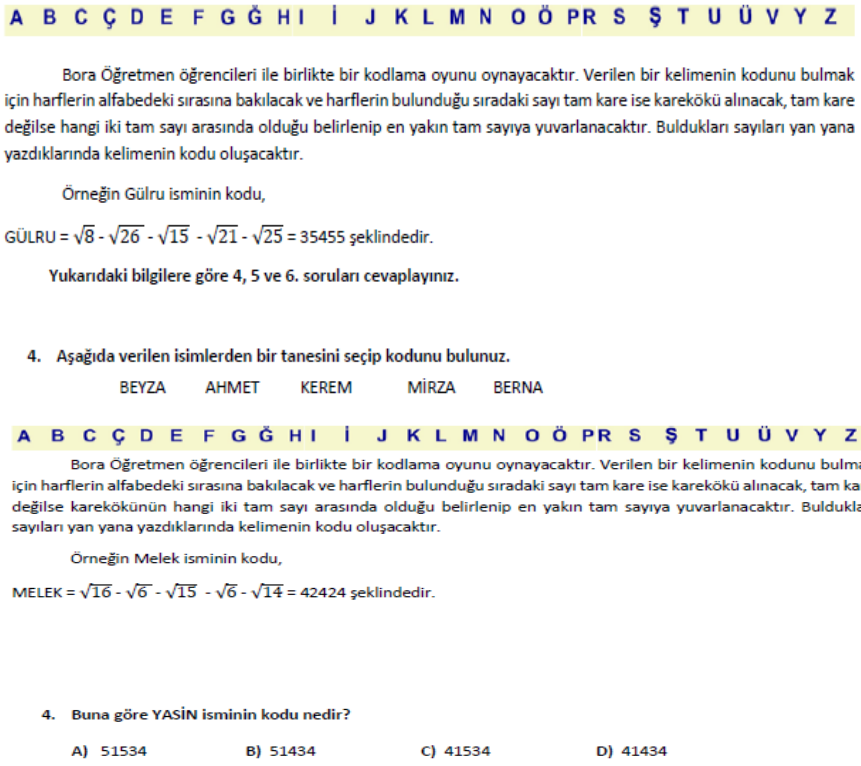
The study group of the research consists of 8th-grade students studying in public elementary schools affiliated with the Ministry of National Education in Ankara province in the 2022-2023 academic year. Ethical approval of the research was obtained from the Gazi University Ethics Commission. The pilot study was conducted in 3 elementary schools in Ankara and Beypazarı district. These schools were selected because of the large number of students they have. Since the number of students in Beypazarı district was insufficient, the final implementation was carried out in a public elementary school in the central district of Ankara. While determining this school, it was taken into consideration that it should be similar to the achievement levels of the schools in the pilot study. In the pilot study, 119 students solved the mathematics achievement test consisting of open-ended items, and 115 students solved the mathematics achievement test consisting of multiple-choice items. Of the 119 students who solved the achievement test consisting of open-ended items, 31 students were not included in the analysis because they either left all the items blank or scored zero points. As a result, 88 students' responses to open-ended items and 115 students' responses to multiple-choice items were analyzed in the pilot application. The final application was carried out with the participation of 157 students. Since it was observed that 29 of these students either did not answer the achievement test consisting of open-ended items at all or answered incorrectly and received zero points, the results of these students were not included. In the final application, the responses of 128 students to the achievement tests consisting of both open-ended and multiple-choice items were evaluated.

Data Collection Tools

The data required for this study were collected through two separate mathematics achievement tests with testlets consisting of open-ended and multiple-choice items. First, two achievement tests with similar content, one with open-ended items and the other with multiple-choice items, were prepared for pilot testing. Each test included four testlets and four items in each testlet. The items were prepared in line with the achievements of "exponential expressions, square root expressions, data analysis, and probability of simple events" from the 8th-grade mathematics curriculum of the 2022-2023 academic year. In the assessment of testlets consisting of open-ended items, rubric and evaluation form were prepared for each item. Accordingly, the grade was calculated by 3 points for the entirely correct answers and 2 points and 1 point for the partially correct answers. Blank and other answers were evaluated as 0 points. An example of open-ended and multiple-choice items from the same content is given in Figure 1.

Figure 1

Example of Open-Ended and Multiple-Choice Test Items



A B C Ç D E F G Ğ H I İ J K L M N O Ö PR S Ş T U Ü V Y Z

Bora Öğretmen öğrencileri ile birlikte bir kodlama oyunu oynayacaktır. Verilen bir kelimenin kodunu bulmak için harflerin alfabedeki sırasına bakılacak ve harflerin bulunduğu sıradaki sayı tam kare ise karekökü alınacak, tam kare değilse hangi iki tam sayı arasında olduğu belirlenip en yakın tam sayıya yuvarlanacaktır. Buldukları sayıları yan yana yazdıklarında kelimenin kodu oluşacaktır.

Örneğin Gülru isminin kodu,

GÜLRU = $\sqrt{8}$ - $\sqrt{26}$ - $\sqrt{15}$ - $\sqrt{21}$ - $\sqrt{25}$ = 35455 şeklindedir.

Yukarıdaki bilgilere göre 4, 5 ve 6. soruları cevaplayınız.

4. Aşağıda verilen isimlerden bir tanesini seçip kodunu bulunuz.

BEYZA AHMET KEREM MİRZA BERNA

A B C Ç D E F G Ğ H I İ J K L M N O Ö PR S Ş T U Ü V Y Z

Bora Öğretmen öğrencileri ile birlikte bir kodlama oyunu oynayacaktır. Verilen bir kelimenin kodunu bulmak için harflerin alfabedeki sırasına bakılacak ve harflerin bulunduğu sıradaki sayı tam kare ise karekökü alınacak, tam kare değilse karekökünün hangi iki tam sayı arasında olduğu belirlenip en yakın tam sayıya yuvarlanacaktır. Buldukları sayıları yan yana yazdıklarında kelimenin kodu oluşacaktır.

Örneğin Melek isminin kodu,

MELEK = $\sqrt{16}$ - $\sqrt{6}$ - $\sqrt{15}$ - $\sqrt{6}$ - $\sqrt{14}$ = 42424 şeklindedir.

4. Buna göre YASİN isminin kodu nedir?

A) 51534 B) 51434 C) 41534 D) 41434

Seven expert opinions were obtained for the pilot forms of the developed open-ended and multiple-choice tests. In this context, three experts were experts in both mathematics and measurement and evaluation areas, one of whom was a faculty member, and the other two were teachers at the graduate level. In addition, two academicians in the area of measurement and evaluation and one undergraduate mathematics teacher were also consulted. In line with expert opinions, revisions were made to the items based on content, form and item writing rules. After the revisions made by considering the expert opinions, Turkish language expert opinion was also taken. The experts chose one of the appropriate options from the expressions "appropriate" or "not appropriate" for each item while expressing their opinions. If the experts chose the same option for the same item, it was considered agreement, and if they chose different options, it was considered disagreement. In this study, inter-expert agreement was calculated to ensure validity and reliability. For this purpose, Miles and Huberman's (1994) reliability formula was used to determine the percentage of agreement between experts. According to the formula, the percentage of agreement is expressed as "Reliability = (Agreement / Agreement + Disagreement) * 100". Accordingly, the percentages of inter-expert agreement were calculated using Miles and Huberman's (1994) formula, and the average percentage of inter-expert agreement was found to be 85%. In order for the research to be considered reliable, the reliability estimates must be above 70% (Miles & Huberman, 1994). Therefore, the result obtained in this study indicates that inter-rater agreement was achieved. Experts were also asked to give their opinions on whether the open-ended and multiple-choice items had similar content. The experts expressed their opinions as "similar content", "partially similar content" and "not similar content". None of the experts chose the "not similar content" option. The percentage of agreement of the expert opinions on the similarity of the content was calculated, and the lowest was 75% and the highest was 100% and the average was 89%. Thus, the pilot application of the tests, which were decided to be appropriate in terms of content, language and expression, was started.

The pilot study aimed to determine the item difficulty and discrimination indices of the open-ended and multiple-choice test items. The open-ended items were scored independently by three mathematics teachers. The raters were given a brief explanation about the measured feature, item content, and rubric

before scoring. The item statistics of the data obtained within the scope of the pilot application are shown in Table 1.

Table 1
Item Statistics of Tests in Pilot Study

		Item statistics for multiple-choice items		Item statistics for open-ended items	
Testlet	Item No	p	r	p	r
1	1	0,50	0,47	0,36	0,32
	2	0,46	0,46	0,33	0,71
	3	0,43	0,38	0,32	0,62
	4	0,33	0,24	0,17	0,40
2	5	0,58	0,62	0,47	0,84
	6	0,63	0,52	0,34	0,67
	7	0,37	0,57	0,09	0,22
	8	0,34	0,37	0,01	0,05
3	9	0,42	0,43	0,52	0,60
	10	0,18	0,19	0,03	0,10
	11	0,15	0,14	0	0
	12	0,33	0,45	0	0
4	13	0,48	0,48	0,14	0,41
	14	0,42	0,60	0,06	022
	15	0,36	0,55	0,08	0,27
	16	0,29	0,53	0,08	0,21

When deciding on the items to be used in the final application, the discrimination (r) of the multiple-choice items was taken into consideration and a selection was made accordingly. Generally, items with an item discrimination between 0.20 and 0.30 are considered usable in the test; items with an item discrimination between 0.30 and 0.40 are considered good; and items with an item discrimination higher than 0.40 are considered very good. It is recommended that items with discrimination lower than 0.20 should be corrected and improved (Özçelik, 2013). Since balanced designs were examined in this study, one item from the testlet was removed. Items with low discrimination (4, 8, 11 and 16) were removed from the test, and item 10 was corrected and included in the test. Since items 11 and 12 of the open-ended items were not responded by any student, the item statistics were zero. It is thought that this situation is caused by the fact that the responses to these items were prepared as free responses. Since the results differed, it was decided to take the opinions of two faculty members in the field of measurement and evaluation and use expert opinions that the items had similar content instead of comparing the item statistics one-to-one.
In line with this purpose, it was decided which items should be removed or revised, and the final achievement test forms were created with four testlets consisting of 3 items each.

Implementation Process

Since the achievement tests included the 8th-grade first-semester subjects, the pilot application was carried out after these subjects were covered. After the pilot application, the necessary analyses were made, and the final tests were created.
For the final application, one group was first administered an achievement test consisting of open-ended items, while the other group was administered an achievement test consisting of multiple-choice items. In this way, it was aimed to prevent an effect caused by the order of administration of tests consisting of different item formats. A ten-day break was given for the administration of the second test. After ten days, the test consisting of multiple-choice items was administered to the group to which the test consisting of open-ended items was administered first, and the test consisting of open-ended items was administered to the group to which the test consisting of multiple-choice items was administered.

Data Analyses

In order to obtain more reliable results in the analyses, three mathematics teachers served as raters. The mathematics teachers conducted their scoring independently. G and Phi coefficients were calculated in the tests consisting of open-ended and multiple-choice testlets within the framework of G theory. In

addition, the D study was conducted by selecting the appropriate variables from the number of items, testlets, and raters within each design. EduG program was used in data analysis.

Results

Results Related to Sub-Problem 1

1.a. In the study, the responses of 128 8th-grade students to an achievement test consisting of 12 open-ended items were analyzed within the framework of G theory without considering the testlet effect. In this context, the results of the G study belonging to the p x i x r design in which person (p), item (i) and raters (r) were crossed with each other are given in Table 2.

Table 2

G Study Results for the p×i×r Design in which the Testlet Effect is not Handled in the Achievement Test with Open-Ended Items

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	1352,284	127	10,648	0,238	19,3
r	7,461	2	3,731	0,000	0,0
i	551,671	11	50,152	0,115	9,3
pr	91,816	254	0,361	0,007	0,6
pi	2773,607	1397	1,985	0,571	46,2
ri	92,247	22	4,193	0,031	2,5
pri,e	763,809	2794	0,273	0,273	22,1
Total	5632,895	4607			100%
G coefficient	0,81				
Phi coefficient	0,78				

When Table 2 is analyzed, the variance belonging to persons (p) explains 19.3% of the total variance and indicates the extent to which persons differ from each other. The value of this variance component is expected to be quite high. It is seen that most of the variability is explained by other sources of variability. The rater (r) variance component (0.000) indicates excellent consistency between the raters' ratings. The item (i) variance component accounts for 9.3% of the total variance and suggests that the difficulty levels of the items differ. The value obtained for the variance component of the person x rater (pr) interaction is 0.007, explaining 0.6% of the total variance. This value means that the scores given by the raters to the persons did not differ much between the raters. In other words, the raters gave similar scores to the persons, which is a desirable situation. When the person × item (pi) interaction variance component is analyzed, it accounts for 46.2% of the total variance and has the highest variance value. This value indicates that the difficulty levels of the items differ from one person to another. The value calculated for the variance component of the rater x item (ri) interaction is 0.031, explaining 2.5% of the total variance. This value indicates the variability of the scores given by the raters to the items. It is desirable that this value is low. The residual component accounts for 22.1% of the total variance, pointing that there are systematic or non-systematic error sources in this study with the interaction between persons, items and raters. Finally, in the analyses obtained as a result of the G theory study, it is seen that the G and Phi coefficient is calculated as 0.81 and 0.78, respectively. Since these values are well above 0.70, they are acceptable values.

1.b. The results of the D study for the p×i×r design in which the testlet effect was not considered in the achievement test consisting of open-ended items are shown in Table 3.

Table 3

D Study Results for the p×i×r Design in which the Testlet Effect is Not Handled in an Achievement Test Consisting of Open-Ended Items

Condition-1	Number of Items	G	Phi	Condition-2	Number of Raters	G	Phi
Number of persons: 128 Number of Raters: 3	6	0,68	0,64	Number of persons: 128 Number of Items: 12	2	0,79	0,76
	9	0,76	0,73		**3***	**0,81**	**0,78**
	12*	**0,81**	**0,78**		4	0,81	0,78
	15	0,84	0,81		5	0,82	0,79
	18	0,86	0,84		6	0,82	0,79
	21	0,88	0,86		7	0,82	0,79

* Refers to the case study data of the current study.

Table 3 shows that when the number of persons and raters is kept constant and the number of items was increased, the coefficients of G and Phi were also increased. When the number of persons and items were kept constant and the number of raters was increased, the reliability coefficients were increased, but they were not affected as much as the increase in the number of items.

1.c. The results of the G study for the p×(i:t)×r design in which the testlet effect was taken into account in the achievement test consisting of open-ended items are shown in Table 4.

Table 4

G Study Results for the p×(i:t)×r Design Considering the Testlet Effect in an Achievement Test Consisting of Open-Ended Items

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	1352,284	127	10,648	0,199	16,0
r	7,461	2	3,731	-0,001	0,0
t	145,362	3	48,454	-0,005	0,0
i:t	406,309	8	50,789	0,119	9,6
pr	91,816	254	0,361	0,003	0,3
pt	1318,138	381	3,460	0,218	17,6
pi:t	1455,469	1016	1,433	0,392	31,6
rt	30,228	6	5,038	0,003	0,2
ri:t	62,019	16	3,876	0,028	2,3
prt	243,605	762	0,320	0,021	1,7
prt:i,e	520,203	2032	0,256	0,256	20,6
Total	5632,895	4607			100%
G Coefficient	0,67				
Phi Coefficient	0,65				

When Table 4 is examined, it is seen that the calculated value of the i:t variance component in which the items are nested in the testlets is 0.119, accounting for 9.6% of the total variance. This value indicates that there is a slight difference between the difficulty levels of the items in the testlets. The variance component of the person-testlet accounts for 17.6% of the total variance and indicates that there are differences due to the person-testlet interaction. The variance component of the person x item interaction within the testlet (pi:t) has the largest value (0.392). This value alone accounts for 31.6% of the total variance and indicates that the person-item interaction varies within the testlet. The G coefficient was 0.67 and the Phi coefficient was 0.65, and it is seen that the G and Phi coefficients are lower than when the testlet effect is not handled.

1.d. The results of the D study for the p×(i:t)×r design in which the testlet effect was taken into account in the achievement test consisting of open-ended items are given in Table 5.

Table 5

D Study Results for the p×(i:t)×r Design Handling the Testlet Effect in an Achievement Test Consisting of Open-Ended Items

Condition-1	Number of Teslets	G	Phi	Condition-2	Number of Items in a Testlet	G	Phi	Condition-3	Number of Raters	G	Phi
Number of persons: 128 Number of raters: 3 Number of items in the testlet: 3	3	0,61	0,58	Number of persons: 128 Number of raters: 3 Number of testlets: 4	1	0,53	0,49	Number of persons: 128 Number of testlets: 4 Number of items in the testlet: 3	2	0,66	0,64
	4*	**0,67**	**0,65**		2	0,63	0,60		**3***	**0,67**	**0,65**
	5	0,72	0,70		**3***	**0,67**	**0,65**		4	0,68	0,65
	6	0,75	0,73		4	0,69	0,67		5	0,68	0,66
	7	0,78	0,76		5	0,71	0,69		6	0,68	0,66
	8	0,80	0,78		6	0,72	0,71		7	0,68	0,66

* Refers to the case study data of the current study.

Table 5 shows that the coefficients of G and Phi increase when the number of testlets, the number of raters and the number of items in the testlets increase, respectively. It is observed that the G coefficient increases above 0.70 with four testlets consisting of five items each (total 20 items) when the number of items in the testlets increases, while it increases above 0.70 with five testlets consisting of three items each (total 15 items) when the number of testlets increases. It can be stated that the increase in the number of raters did not affect the reliability coefficients to a great extent.

Results Related to Sub-Problem 2

2.a. The results of the G study for the *p×i* design in which the testlet effect was not handled in the achievement test consisting of multiple-choice items are shown in Table 6.

Table 6

G Study Results for the p×i Design in Achievement Test Consisting of Multiple-Choice Items in which the Testlet Effect is not Handled

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	100,093	127	0,788	0,050	20,2
i	19,898	11	1,809	0,013	5,1
pi,e	258,852	1397	0,185	0,185	74,7
Total	378,843	1535			100%
G Coefficient	0,76				
Phi Coefficient	0,75				

Table 6 demonstrates that the estimated variance component for the individuals accounts for 20.2% of the total variance, with a value of 0.050. The fact that this variance component belonging to persons is quite high indicates that there is a systematic difference between persons and this is an expected situation. With a value of 0.013, the item variance component's calculated value accounts for 5.1% of the total variance. Here, it is possible to say that item difficulties do not differ much. The highest variance value belongs to the residual component with 0.185. This value accounts for 74.7% of the total variance. This is an indication that the person-item interaction and systematic or non-systematic error sources that could not be measured in this study were not controlled. This variance belonging to the residual component is expected to be quite low. The G coefficient was 0.76 and the Phi coefficient was 0.75, and it can be stated that the reliability coefficients are at an adequate level.

2.b. The results of the D study for the *p×i* design in which the testlet effect was not handled in the achievement test consisting of multiple-choice items are shown in Table 7.

Table 7
D Study Results for the p×i Design in which the Testlet Effect is Not Handled in an Achievement Test Consisting of Multiple-Choice Items

Condition-1	Number of Items	G	Phi
Number of persons: 128	6	0,62	0,60
	9	0,71	0,70
	12*	0,76*	0,75*
	15	0,80	0,79
	18	0,83	0,82
	21	0,85	0,84

* Refers to the case study data of the current study.

When Table 7 is examined, it is observed that in achievement tests consisting of multiple-choice items, the G and Phi reliability coefficients obtained in the p×i design in which persons (*p*) are crossed with items (*i*) increase as the number of items increases. When the number of items increased from 6 to 15, the G coefficient increased from 0.62 to 0.80.

2.c. The results of the G study for the p×(i:t) design in which the testlet effect is taken into account in the achievement test consisting of multiple-choice items are presented in Table 8.

Table 8
G Study Results for the p×(i:t) Design Handling the Testlet Effect in an Achievement Test Consisting of Multiple-Choice Items

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	100,093	127	0,788	0,047	18,8
t	7,929	3	2,643	0,003	1,1
i:t	11,969	8	1,496	0,010	4,2
pt	86,154	381	0,226	0,019	7,5
pi:t,e	172,698	1016	0,170	0,170	68,3
Total	378,843	1535			100%
G Coefficient	0,71				
Phi Coefficient	0,70				

When Table 8 is analyzed, the variance belonging to the main effect of persons (*p*) accounts for 18.8% of the total variance. This value is the second largest percentage of variance in the table 8, indicating that there is a systematic difference between persons. With a value of 0.003, the variance estimated for the testlet (t) main effect explains 1.1% of the overall variance. This value means that the difficulty levels of the testlets do not differ from each other. With a value of 0.010, the variance component of the i:t effect, which involves items nested within testlets, accounts for 4.2% of the total variance. The fact that this value is close to zero indicates that the difficulty levels of the items in the same testlet are close to each other. The estimated variance of person testlet (pt) accounts for 7.5% of the total variance. Here, it can be stated that there are differences due to person-testlet interaction. In this design, the largest variance value belongs to the residual component (pi:t,e), accounting for 68.3% of the total variance. This value is lower than the accounted percentage of the residual component (74.7%) obtained when the testlet effect is not handled. This value shows that since there are more variance sources when the testlet effect is handled, the percentage of accounted total variance is divided into these variance sources and the percentage of variance belonging to the residual component decreases. When the reliability coefficients are analyzed, it is seen that the G coefficient is 0.71 and the Phi coefficient is 0.70.

2.d. The results of the D study for the p×(i:t) design in which the testlet effect is taken into account in the achievement test consisting of multiple-choice items are given in Table 9.

Table 9

D Study Results for the p×(i:t) Design Handling the Testlet Effect in an Achievement Test Consisting of Multiple-Choice Items

Condition-1	Number of Testlets	G	Phi	Condition-2	Number of Items in a Testlet	G	Phi
.Number of persons: 128 Number of items in a testlet: 3	2	0,55	0,53	Number of persons: 128 Number of Testlets: 4	2	0,64	0,63
	3	0,65	0,63		**3***	**0,71**	**0,70**
	4*	**0,71**	**0,70**		4	0,75	0,74
	5	0,76	0,74		5	0,78	0,76
	6	0,79	0,77		6	0,80	0,78
	7	0,81	0,80		7	0,81	0,80

* Refers to the case study data of the current study.

When Table 9 is analyzed, it is seen that the G and Phi coefficients increase when the number of persons and items in the testlets are kept constant and the number of testlets is increased. In the second case, the reliability coefficients increased when the number of individuals and testlets were kept constant and the number of items in the testlet was increased. Reliability coefficients are found to be more impacted by an increase in testlet count than by an increase in testlet item count. This result is similar to the results obtained with testlets consisting of open-ended items. Obtaining the same G coefficient with more items indicates that the increase in testlets is more effective.

Conclusion and Discussion

When the findings of the achievement tests consisting of open-ended items were analyzed, the reliability coefficients that were estimated differed when the testlet effect was not taken into account and when it was taken into account. This result can be stated as an expected situation in testlet consisting of open-ended items with low objective scoring (Kaya Uyanık & Ertuna, 2022). In the case where the testlet effect is not handled, it overestimates the reliability value of the test due to the correlation between the items in the testlet (Lee & Park, 2012). As a result of the D studies, the reliability coefficients were increased when the number of items increased by keeping the number of persons and raters constant in achievement tests consisting of open-ended items. The increase in the number of items also increases the reliability of the test (Baykul, 2000; Turgut, 1992). However, it was observed that the reliability coefficients increased to a certain number of items, and after a certain number of items, the increase did not contribute as much as before. For this reason, it should first be decided whether the test to be applied will be high-stakes testing or low-stakes testing. If it is a high-stakes test, it is recommended that the reliability should be 0.80 and above (Nunnally, 1967; as cited in Henson, 2001). Then, choosing the number of items according to the reliability coefficients at the level that will serve the test's purpose would be appropriate. In tests consisting of open-ended items, it was discovered that adding more testlets to the test was more beneficial than adding more items to the testlets. When the increase in the number of testlets is compared with the increase in the number of items, it can be stated that the increase in the number of testlet has a higher contribution to the increase in the number of items in the test. One of the reasons the rater variance was very low in the analyses according to both different designs in the tests with open-ended items is the use of a detailed rubric. Since the use of rubrics can increase objectivity (Moskal & Leydens, 2000), it may have reduced the error caused by the rater. The decision study on the number of raters determined that the increase in the number of raters did not have much effect on the reliability value. The study concluded that two raters would be sufficient due to the difficulty and inconvenience of finding a large number of raters and in terms of time and practicality. This result is similar to the results of Kaya Uyanık and Ertuna (2022). Taşdelen Teker et al (2016) obtained sufficient reliability value with two raters as a result of the D study conducted in their study in which students' communication skills were evaluated with a 5-point rating scale.

In the test consisting of multiple choice items, as in the test with open-ended ones, similar results were obtained: the reliability coefficients were estimated higher when the testlet effect was not taken into account, and the reliability coefficients were estimated lower when the testlet effect was taken into account. This result was similar with the results of studies examining the testlet consisting of multiple choice items and its reliability effect (Hendrickson, 2001; Lee & Park, 2012; Sireci et al, 1991; Taşdelen Teker, 2014; Thissen et al, 1989; Wainer, 1995). Situations, where the testlet effect is not handled may cause bias in the results and a higher estimate of the reliability value. A high correlation between items within the same testlet will also contribute to homogeneity. If the testlet effect is taken into account, the different contents of the testlets will provide heterogeneity. This may lead to a lower estimate of reliability when the testlet effect is taken into account than when it is not taken into account. It is seen in the D studies that when the number of individuals is kept constant, reliability will increase more when testlets are increased rather than when items are increased. In their simulation studies, Kaya Uyanık and Gelbal (2018) obtained a higher reliability value when the number of items increased if the testlets were equal, similar to the results of this study. In the event that each testlet had the same number of items, higher reliability was obtained as the number of testlets increased. In short, higher reliability is achieved when the total number of items increases.

In their study with dummy-coded SAT data, Sireci et al.'s (1991) determined that not taking into account the relationship between items in the same testlet led to a 10-15% overestimation in both the CTT-based and the IRT-based reliability estimation. In this study, higher G and Phi coefficients were obtained by ignoring the testlet effect in both the test consisting of open-ended items and the test consisting of multiple-choice items. If the testlet effect was taken into account in the test with open-ended items, it caused a decrease in the G coefficient (difference of 0.14) and Phi coefficient (difference of 0.12). The similar difference is greater than the difference in G coefficient (difference of 0.05) and Phi coefficient (difference of 0.05) in the achievement test, which includes testlets of multiple-choice items. Therefore, it can be stated that the item types are effective in the change in G and Phi coefficients.

Within the scope of the research, an achievement test for mathematics courses was developed. The results of the research in different fields can be examined. Since the number of items in the testlets was equal in this research, the studies were conducted on balanced designs. Research can be conducted in unbalanced designs where the number of items in the testlets varies. In addition, the results can be examined by conducting studies with different designs in which raters, which are not included in the scope of this research, are nested within persons or items are nested within raters.

Declarations

Author Contribution: Serpil KOCAOĞLU: conceptualization, methodology, analysis, writing & editing, visualization. Melek Gülşah ŞAHİN: conceptualization, methodology, writing-review & editing, supervision.

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

Ethical Approval: The study was ethically approved by the Gazi University Ethics Commission dated 30.12.2022 and numbered E.548756.

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

References

Attali, Y., Laitusis, C., & Stone, E. (2016). Differences in reaction to immediate feedback and opportunity to revise answers for multiple-choice and open-ended questions. *Educational and Psychological Measurement, 76*(5), 787-802. https://journals.sagepub.com/doi/10.1177/0013164415612548

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* ÖSYM.

Berberoğlu, G. (2006). *Sınıf içi ölçme değerlendirme teknikleri*. Morpa Kültür.

Brennan, R. L. (2001). *Generalizability theory.* Springer-Verlag.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271. https://doi.org/10.2307/145138

Doğan, N. (2009). Yazılı yoklamalar. *In H. Atılgan (Ed.), Eğitimde ölçme ve değerlendirme* (p.148). Anı.

Doğan, N. (2019a). Geleneksel ölçme ve değerlendirme teknikleri I: Yanıtı seçmeyi gerektiren ölçme araçları. *In N. Doğan (Ed.), Eğitimde ölçme ve değerlendirme* (pp. 113-138). Pegem Akademi.

Doğan, N. (2019b). Geleneksel ölçme ve değerlendirme teknikleri II: Yanıtı yapılandırmayı gerektiren ölçme araçları. *In N. Doğan (Ed.), Eğitimde ölçme ve değerlendirme* (pp:140-179). Pegem Akademi.

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

Gessaroli, M. E., & Folske, J.C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing, 2*(3-4), 277-295. https://doi.org/10.1080/15305058.2002.9669496

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı.* Pegem Akademi.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Taylor & Francis Group. https://ebookcentral.proquest.com/lib/gazi-ebooks/detail.action?docID=255610

Hendrickson, A. B. (2001). *Reliability of scores from tests composed of testlets: A comparison of methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177-189. https://www.tandfonline.com/doi/abs/10.1080/07481756.2002.12069034

Karasar, N. (1994). *Bilimsel Araştırma Yöntemi.* 3A Araştırma Eğitim Danışmanlık.

Karatoprak Erşen, R., & Gündüz, T. (2023). Seçme ve katkı gerektiren maddelerin yazımı ve düzenlenmesi için kontrol listeleri. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi* (58), 2473-2493. https://doi.org/10.53444/deubefd.1279240

Kaya Uyanık, G., & Ertuna, L. (2022). Examination of testlet effect in open-ended items. *SAGE Open*, 1-12. https://doi.org/10.1177/21582440221079849

Kaya Uyanık, G., & Gelbal, S. (2018). Madde tepki modellemesinde genellenebilirlik ile iki yüzeyli desenlerin incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 9*(1), 17-32. https://doi.org/10.21031/epod.349718

Ko, M. H. (2010). A comparision of reading comprehension tests: Multiple-choice vs. open-ended. *English Teaching, 65*(1), 137-159. doi:10.15858/engtea.65.1.201003.137

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*(3), 237–255. https://doi.org/10.1207/S15324818AME1203_2

Lee, G., & Park, I.-Y. (2012). A comparison of the approaches of generalizability theory and item response theory in estimating the reliability of test scores for testlet-composed tests. *Asia Pacific Education Review, 13*(1), 47-54. https://doi.org/10.1007/s12564-011-9170-0

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. Educational Measurement: Issues and Pratice, 19(4), 9-15. https://doi.org/10.1111/j.1745-3992.2000.tb00041.x

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation, 7*(10), 1-6. https://doi.org/10.7275/q7rm-gg74

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analyses: An expanded sourcebook.* CA: Sage Publications.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessments of students* (6th ed.). Essex: Pearson International.

Özçelik, D.A. (2013). *Test hazırlama kılavuzu*. Pegem Akademi.

Popham, J.W. (2014). Selected-response tests. *In Classroom assessment: What teachers need to know* (7th ed, pp. 155-180). Pearson Education Ltd.

Russell, M. & Airasian, P.(2008). Designing, administering, and scoring achievement tests. *Classroom assessment: Concepts and applications* içinde (7th ed, pp. 144-175). McGrawHill Higher Education.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Sage Publicatons.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliabilty of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Taşdelen Teker, G. (2014). *Madde takımlarının güvenirlik ve değişen madde fonksiyonu üzerine etkisi.* Doctoral Dissertation, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Taşdelen Teker, G., Şahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment. *Journal of Human Sciences, 13*(3), 5574-5586. https://doi.org/10.14687/jhs.v13i3.4155

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme.* Yargı.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Turgut, M. F. (1992). *Eğitimde ölçme ve değerlendirme metotları.* Saydam.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8*(2), 157-186. https://doi.org/10.1207/s15324818ame0802_4

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice, *15*(1), 22-29. http://doi: 10.1111/j.1745-3992.1996tb00803.x

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. *In W. J. van der Linden & G. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice* (pp. 245–269). Springer. https://doi.org/10.1007/0-306-47531-6_13

Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri. *Gazi Eğitim Bilimleri Dergisi, 2*(2), 151-170. https://dergipark.org.tr/tr/pub/gebd/issue/35205/390659