Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *Kış* 2024 *Winter* 2024 *Cilt: 15-Sayı: 4* *Volume: 15-Issue: 4*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

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

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

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

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

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

Hakem Kurulu / Referee Board

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

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2024, Sayı: 15-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2024, Issue: 15-4)

İÇİNDEKİLER / CONTENTS

Effect Of Content Balancing on Measurement Precision in Computer Adaptive Testing Applications
İlkay ÜÇGÜL ÖCAL, Nuri DOĞAN 395

The Effect of Missing Data Handling Methods on Differential Item Functioning with Testlet Data
Rabia AKCAN, Kübra ATALAY KABASAKAL.....408

Natural Language Processing and Machine Learning Applications For Assessment and Evaluation in Education: Opportunities and New Approaches
Kübra YILMAZ, Kaan Zülfikar DENİZ.....421

FAfA: Factor Analysis for All An R Package to Conduct Factor Analysis with R Shiny Application
Abdullah Faruk KILIÇ.....446

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 395-407

Effect Of Content Balancing on Measurement Precision in Computer Adaptive Testing Applications

İlkay ÜÇGÜL ÖCAL * Nuri DOĞAN **

Abstract

This study aims to investigate the effect of content balancing, which involves equal and different weighting of content areas in dichotomous items in computerized adaptive testing (CAT), on measurement precision under different measurement conditions. Conducted as a simulation study, small sample sizes were set at 250, while large sample sizes comprised 500 individuals. The ability parameters of the individuals forming the sample were generated to display a normal distribution within the range of -3 to +3 for each sample. Using the three-parameter logistic (3PL) item response model, a pool of 750 items spanning five different content areas was developed for dichotomous items. The study considered different sample sizes, ability estimation methods (Maximum Likelihood Estimation and Expected A Posteriori), and termination rules (20 items, 60 items, and SE≤.30) as significant factors in the CAT algorithm for examining the effect of content balancing. For each CAT application, measurement precision was assessed by calculating the root mean square error (RMSE), bias, and fidelity coefficients, and these were analyzed comparatively. The results showed that bias values were close to zero under all conditions. RMSE values were lowest when the test was terminated at 60 items across all conditions, while standard error termination rules and situations where the test terminated at 20 items produced similar values. Considering all conditions, the highest fidelity coefficient was observed when the test terminated at 60 items. The fidelity coefficient did not vary significantly with other variables. Implementing content balancing in conditions using different ability estimation methods increased the average number of items by approximately one item. While the average number of items in the test slightly increased with content balancing, measurement precision was maintained. Overall, the maximum item exposure rate decreased with content balancing when content areas were weighted equally, whereas it increased when they were weighted disproportionately.

Keywords*: computerized adaptive testing, content balancing, measurement precision.*

Introduction

Examinations used in education have traditionally focused on paper and pencil tests and performance assessments. Since the late 1980s, with the widespread adoption of personal computers in education, these examinations have rapidly expanded into formats suitable for computer delivery (Şenel, 2021; Van der Linden & Glas, 2002). Computerized adaptive tests (CATs) utilise an algorithmic approach to administer test items. Specifically, the items selected and administered are tailored to the estimated ability level of the examinee during the testing process, with the estimated ability continually updated after each item is administered. Therefore, CAT is an adaptable test at the item level and can be of fixed or variable length. Ability estimation is used not only to represent an examinee's level of ability but also to determine the selection of subsequent items from the available item pool. CATs can be considerably more useful and efficient than traditional linear tests, which has led to their widespread use in recent years (Cheng & Chang, 2007; Kalender, 2009). Several advantages of CATs over traditional linear tests have been demonstrated, including increased flexibility in test administration, elimination of the need for answer sheets and trained test administrators, enhanced test security, and the ability to provide accurate measurements across a wide range of ability levels (Rudner, 1998; Tian et al., 2007).

* Measurement and Evaluation Specialist, Ministry of National Education, Ankara, Türkiye, ilkayocal83@gmail.com, ORCID ID: 0009-0004-2246-6909
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Üçgül Öcal, İ., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications*, Journal of Measurement and Evaluation in Education and Psychology*, *15*(4), 394-407. https://doi.org/10.21031/epod.1438977

Received: 18.02.2024
Accepted: 18.12.2024

The mathematical model used for CAT applications is based on Item Response Theory (IRT). IRT methodologies are employed in various CAT processes and focus on improving the accuracy and efficiency of ability estimation. IRT-based CAT applications typically contain fewer items than traditional paper and pencil measurements (Embretson & Reise, 2000). The CAT process requires a calibrated item pool and is implemented in four consecutive steps (Thompson & Weiss, 2019):

1. The initial step involves selecting one or several items to start the CAT.
2. The testing step, where items are selected iteratively and optimally, is administered, and ability estimation is performed after each item administration.
3. The termination step defines rules for stopping the adaptive item administration.
4. The final step involves final ability estimation and reporting.

The initial step involves selecting the first item(s) to be administered in the CAT. A commonly used starting rule is the selection of an item that corresponds to the average ability level of the examinee group (theta=0). If no information is available about the examinees' ability levels at the start, this method is considered appropriate. An alternative entry rule could be the selection of an item of medium difficulty ($-0.5<b<0.5$) at the start. After administering the initial item, the cycle of ability estimation and item administration continues until the testing process concludes. Various estimation methods are available for ability estimation. The most commonly used methods include Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP). When item parameters are known, ability parameters can be simply estimated using the maximum likelihood estimation method. This method has several advantages, including consistency and asymptotic normality. For the MLE estimation method to be applicable, the response pattern must contain at least one correct and one incorrect answer. In cases where all items are answered correctly or all are answered incorrectly, the use of the MLE estimation method is not appropriate. In such cases, Bayes-based ability estimation methods, such as EAP or MAP, can be used to overcome this problem. Bayes-based estimation methods have smaller standard errors compared to MLE but require prior knowledge of the individual's ability. The choice of which ability estimation method to use should be made considering all components of the CAT application (Hambleton & Swaminathan, 1985). In the testing step, a hybrid rule that starts with one estimation method and then switches to another after a certain number of items or under certain conditions can also be preferred (Magis et al., 2017). Item selection is a critical component of CAT applications. After determining that test items are appropriate based on the content characteristics in the content balancing component of the CAT algorithm, these items are considered for selection as the next item to be administered. A comprehensive range of item selection methods has been developed in the testing measurement field, yet very few of these methods are employed in actual CAT applications (Han, 2018). One of the best-known and oldest item selection methods is the Maximum Fisher Information (MFI) method. This method involves selecting an item that has the MFI at a certain θ based on the test items previously administered to the examinee.

Test developers have found that the choice of termination rule is largely dependent on the test purpose, item pool characteristics, and operational constraints (Segall, 2005). The termination rule defines parameters for stopping the adaptive item administration. In general, four main termination rules are identified: (a) length criterion, (b) precision criterion, (c) classification criterion, and (d) information criterion (Van der Linden & Glas, 2002).

Validity is one of the most crucial characteristics sought in tests used in education and psychology. Validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA & NCME, 2014). The constructs measured in educational tests are combinations of different subject and content areas. Ensuring the content validity of a test, representing these subjects and content areas adequately within the test is possible. Depending on the test requirements, item selection in CAT applications must meet the requirements of the defined scope to have a balanced content representation; that is, the CAT application must balance items from each subject area according to predetermined percentages. In individualised tests, different items are administered to examinees, but the same item distribution according to the content area should be provided to each. To obtain valid measurements, there must be a balance between the measured content areas or subject areas. Several content balancing methods have been developed to ensure that CAT maintains the desired distribution of content areas throughout the test. Among the most widely used

methods are the Constrained Computerized Adaptive Testing (CCAT), the Modified Multinomial Model (MMM), and the Modified CCAT (MCCAT).

The CCAT method, proposed by Kingsbury and Zara (1989), is a straightforward and understandable two-stage content balancing control mechanism. The content balancing algorithm selects the most suitable item from a content area with the current item usage frequency rate below the targeted application percentage. The selection of the most suitable item is limited by the item usage frequency rate and content area determined to be below the target percentage for the test. Content areas can be weighted equally or differently according to the structure of the respective course. In this method, at each step of the CAT process, experimental percentage rates for each category are calculated. Subsequently, the category with the greatest difference between the theoretical and experimental values is identified, and the next item is selected from this subgroup before returning to the first step. Based on this method, any desired content distribution can be met if the number of items in each content area in the item pool is sufficiently large to construct the target test. The MMM, as described by Chen and Ankenmann (2004), begins by constructing a cumulative distribution based on the target exposure rates of all content areas. A random number from a uniform distribution is used to select the next content area. When a content area reaches its target percentage, a new multinomial distribution is created using the remaining content areas. This method avoids the highly predictable sequence of content areas seen in the CCAT and ensures that target percentages are met exactly. The MCCAT method, proposed by Leung et al. (2000), modifies the original CCAT by selecting items from any unfulfilled content area rather than the one furthest below its target. This approach helps avoid potential undesirable order effects of the CCAT, ensuring a more balanced and less predictable item selection process.

Decisions made based on the measurement results obtained from CAT applications have significant impacts on all educational stakeholders. Therefore, it is crucial to make valid and reliable estimations with CAT applications. The lack of content comparability can pose a threat to the content validity of scores. Whether or not to balance the content of items administered to examinees is one of the fundamental issues to be addressed when developing a CAT application.

Previous studies have extensively explored various aspects of content balancing in CAT. Cheng and Chang (2007) investigated a two-phase item selection procedure that adapts to content requirements while optimizing item selection, highlighting the impact of flexible content balancing on measurement precision and efficiency. Leung et al. (2000) introduced the MCCAT method, which eliminates the predictability of content sequencing while maintaining balance. In subsequent studies, Leung et al. (2003a, 2003b) examined the multistage a-stratified design (ASTR) combined with content balancing methods like MCCAT and the MMM, demonstrating the effectiveness of these methods in reducing item-overlap rates and enhancing item pool utilization without compromising measurement accuracy. Furthermore, Özdemir and Gelbal (2015) and Sari and Manley (2017) explored the practical applications of content balancing in educational settings, emphasizing its role in maintaining test reliability and validity. Demir (2019) analyzed the effects of content balancing on the precision and fairness of CAT applications, providing insights into the psychometric properties affected by different balancing algorithms. Şahin and Özbaşı (2017) reviewed various content balancing methods, offering a comprehensive overview of the current state of research and practical implications. Additionally, Song (2010) focused on the implementation challenges and solutions for content balancing in large-scale adaptive testing programs, while Yasuda and Hull (2021) demonstrated the application of content balancing in the development of CAT-based versions of specific inventories, showing that it can be implemented without compromising accuracy. However, these aforementioned studies often focused on specific methods or conditions, leaving a gap in understanding the comprehensive effects of content balancing across diverse testing scenarios.

Our study addresses this gap by conducting a detailed simulation analysis of content balancing's impact on measurement precision under varying conditions, including different termination rules, sample sizes, and ability estimation methods. This study seeks to answer the following questions:

1. In computerized adaptive testing applications, when content balancing is not performed; how do measurement precision and ability estimations change according to

- Termination rules (20 items, 60 items, SE≤.30),
- Sample sizes (N=250, N=500),

- Ability estimation methods (MLE, EAP)?

2. In computerized adaptive testing applications, when content areas are weighted equally for content balancing; how do measurement precision and ability estimations change according to

- Termination rules (20 items, 60 items, SE≤.30),
- Sample sizes (N=250, N=500),
- Ability estimation methods (MLE, EAP)?

3. In computerized adaptive testing applications, when content areas are weighted disproportionately for content balancing; how do measurement precision and ability estimations change according to

- Termination rules (20 items, 60 items, SE≤.30),
- Sample sizes (N=250, N=500),
- Ability estimation methods (MLE, EAP)?

Methods

Research Model

This study aims to examine how content balancing in CAT applications with dichotomous items affects measurement precision under different conditions. The nature of this research is descriptive and simulative.

Data Generation

Participants for the CAT application were simulated using the R Studio program by the researcher (R Core Team, 2013). Initially, ability parameter values (true θ) for individuals were obtained, followed by item parameter values. Samples of two different sizes, 250 and 500 individuals, were created. The ability parameters of the individuals taking the test were generated to display a normal distribution θ~N(0, 1) within the range of -3 to +3 for each sample size condition.

The item pool for the CAT applications was created according to the 3PLM using the R Studio program. The item parameters were determined by the researcher to follow a uniform distribution. Feinberg and Rubright (2016) noted that item parameters are often simulated to follow a uniform distribution when using the three-parameter logistic model. For content balancing, item pools consisting of 750 items from five different content areas were created, weighted equally and disproportionately, using the 3PLM. In the item pool where content areas were weighted equally, each content area consisted of 150 items. In the item pool where content areas were weighted disproportionately, the different content areas contained 50, 50, 150, 250, and 250 items, respectively.

CAT applications yield better results when the items in the item pool have a sufficient number and a uniform distribution that caters to different ability levels and when the items are highly discriminative (DeMars, 2010; Flaugher, 2000). Therefore, item discrimination parameters “a” (ranging from 0.5 to 2), item difficulty parameters “b” (ranging from -3 to 3) and guessing parameters “c” (.05 to .2) were generated to follow a uniform distribution (Ree & Jensen, 1983; Thompson, 2009).

CAT Conditions

When no prior information about an individual's ability is available, assuming an average ability level is the most appropriate estimate. Starting the CAT application with an item of average difficulty level will be more psychometrically effective (Mills & Stocking, 1996). Therefore, the method within the range -.50<b<.50 was used as the test initiation rule for the simulative CAT application.

One of the best-known and oldest item selection methods, Maximum Fisher Information (MFI), involves selecting and administering an item that has the maximum Fisher information at a certain condition based on the test items previously administered (Han, 2018; Kalender, 2009). The MFI item selection method was chosen as a fixed condition in the simulation study. In the literature, there are various ability estimation methods based on dichotomous items and unidimensional IRT. The most frequently used among these methods are the MLE method and the Bayesian estimation method EAP (Chen et al., 1998; Segall, 2005). These two methods were considered as conditions for ability estimation in the current study. Fixed-length (20 and 60 items) and ability level's standard error (SE≤.30) rules were determined as conditions for test termination. To observe the performance of content balancing in short and long tests and to ensure adequate representation of all content areas, fixed test lengths of 20 and 60 items

have been chosen. Among the methods proposed for content balancing while maintaining test efficiency, the most frequently used, simple, and understandable method is the CCAT method (Kingsbury & Zara, 1989). In the current simulation study, the CCAT method available in the "catR" package used for data analysis was employed as the content balancing method, leaving other content balancing methods outside the scope of this study. No item exposure rate control was conducted in the CAT application. The CAT conditions determined within the scope of the study are provided in Table 1.

Table 1
Conditions for the Computerized Adaptive Testing Application

CAT Components	Conditions	Number of Conditions
Termination Rule	20 items 60 items SE≤.30	3
Sample Size	250 500	2
Ability Estimation Method	MLE EAP	2
Test Initiation Rule	-.50<b<.50	1
Item Selection Method	MFI	1
Item Exposure Control	None	1
Content Balancing	None Equally Weighted Contents Differentially Weighted Contents	3

In the study, a total of 36 simulation conditions were examined, encompassing 3 termination rules, 2 sample sizes, 2 ability estimation methods, and 3 content balancing scenarios.

Data Analysis

In the scope of the research, measurement precision for each condition was evaluated using fidelity coefficient, RMSE (Root Mean Squared Error), and bias values. For most IRT studies, Harwell et al. (1996) recommended at least 25 replications to reduce sample bias and obtain stable and highly reliable results, but they also noted that in some studies this number may be much higher. These values were calculated separately for each of the 50 replications and then averaged.
The fidelity coefficient was assessed by calculating the correlation between the true θ levels, which were simulated at the start for individuals, and the θ levels estimated in each research condition and replication. The average correlation of the estimated θ values for each participant was obtained by averaging these correlations. The Pearson's correlation coefficient was used to calculate the fidelity coefficient, which is computed using the following formula:

$$r = \frac{\mathrm{cov}(\hat{\theta}, \theta)}{ss(\hat{\theta})ss(\theta)}$$

RMSE, the square root of the average of the squared differences between the estimated parameter value for each item in each replication and the true parameter value, is one of the most commonly used measures to evaluate the accuracy of estimates. It shows how far the estimates deviate from the true values using the Euclidean distance. Bias, indicating the systematic error related to the estimate, is equal to the difference between the average of the estimated parameter values for each item in each replication and the true parameter value, and is another measure indicating the precision of measurement. RMSE and bias values are calculated using the following equations (Zheng & Chang, 2014):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)^2}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)$$

In the equations, n represents the number of individuals, θi represents the individual's true ability level, and θi represents the estimated ability level of the individual. A high fidelity coefficient and low values of bias and RMSE indicate that there is no difference between the true ability level and the estimated ability level. The average test length in conditions where the termination criterion was set as SE≤.30 has also been examined.

To provide insights into test security, the maximum item exposure rates (r_{max}) for each condition were also examined.

Results

In this study, the RMSE, bias, and fit values calculated as indicators of measurement precision under 36 different conditions, along with the average test length in conditions where SE≤.30, are provided in Table 2.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 395-407

Tablo 2

The Impact of Content Balancing on Measurement Precision Under Different Measurement Conditions in Computerized Individualised Testing Applications

Sample Size	Ability Estimation Method	Termination Rule	Content Balancing											
			None				Equally Weighted				Differentially Weighted			
			RMSE	*Bias*	*Correlation*	*Average Number of Items*	*RMSE*	*Bias*	*Correlation*	*Average Number of Items*	*RMSE*	*Bias*	*Correlation*	*Average Number of Items*
250	EAP	20 items	0.1900	0.0368	0.9830	-	0.1935	0.0322	0.9821	-	0.1954	0.0408	0.9795	
250	EAP	60 items	0.1282	0.0463	0.9931	-	0.1283	0.0462	0.9930	-	0.1270	0.0470	0.9922	
250	EAP	SE≤0.30	0.2025	0.0346	0.9805	17.40	0.2028	0.0299	0.9802	18.00	0.2020	0.0386	0.9778	17.97
250	MLE	20 items	0.2044	0.0357	0.9806	-	0.2135	0.0394	0.9791	-	0.2121	0.0486	0.9776	
250	MLE	60 items	0.1327	0.0453	0.9927	-	0.1349	0.0463	0.9925	-	0.1327	0.0494	0.9919	
250	MLE	SE≤0.30	0.2045	0.0348	0.9804	19.01	0.2070	0.0407	0.9801	19.73	0.2066	0.0453	0.9785	19.60
500	EAP	20 items	0.1868	0.0313	0.9821	-	0.1928	0.0325	0.9825	-	0.1962	0.0392	0.9797	
500	EAP	60 items	0.1255	0.0459	0.9929	-	0.1281	0.0453	0.9931	-	0.1287	0.0486	0.9923	
500	EAP	SE≤0.30	0.1968	0.0289	0.9799	17.32	0.2021	0.0314	0.9807	18.02	0.2011	0.0343	0.9783	18.14
500	MLE	20 items	0.2068	0.0385	0.9805	-	0.2132	0.0416	0.9796	-	0.2126	0.0503	0.9780	
500	MLE	60 items	0.1315	0.0465	0.9929	-	0.1329	0.0474	0.9928	-	0.1344	0.0537	0.9921	
500	MLE	SE≤0.30	0.2067	0.0374	0.9804	18.94	0.2097	0.0390	0.9800	19.67	0.2082	0.0464	0.9785	19.75

* Measurement and Evaluation Specialist, Ministry of National Education, Ankara, Türkiye, ilkayocal83@gmail.com, ORCID ID: 0009-0004-2246-6909
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Üçgül Öcal, İ., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications*, Journal of Measurement and Evaluation in Education and Psychology*, *15*(4), 394-407. https://doi.org/10.21031/epod.1438977

Received: 18.02.2024
Accepted: 18.12.2024

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 395-407

For different sample sizes, the estimated ability levels obtained by applying two different ability estimation methods were compared with the individuals' true ability levels in terms of RMSE and bias values.

When it comes to Table 2, bias values were close to zero in all conditions. The highest bias values (0.05) for both sample sizes were obtained when 60 items were used as the test termination rule in the EAP ability estimation method. When the test was terminated at 20 items and with the standard error termination rule, bias values (0.03) were found to be quite close to each other. Similarly, in the MLE ability estimation method, the highest bias value (0.05) for both sample sizes was obtained when 60 items were used as the test termination rule, both when content balancing was not performed and when content areas were equally weighted. When content balancing was performed with differentially weighted content areas, all bias values were relatively high (0.05) compared to other conditions. When the test was terminated at 20 items and with the standard error termination rule, bias values (0.04) were found to be quite close to each other. Generally, bias values were slightly higher in the MLE estimation method compared to the EAP method. Content balancing with equally weighted content areas did not affect bias values in both estimation methods when all conditions were considered together. Additionally, it was observed that bias values slightly increased in conditions of content balancing with differentially weighted content areas.

RMSE values were lowest when the test terminated at 60 items across all conditions, while they were similar for the standard error termination rule and when the test terminated at 20 items. Using the EAP ability estimation method, the lowest RMSE value (0.13) for both sample sizes was obtained when the test terminated at 60 items. When the test terminated at 20 items (0.19) and with the standard error termination rule (0.20), RMSE values were quite close to each other. Content balancing with equally and differentially weighted contents did not cause a significant change in RMSE values. Similarly, when using the MLE ability estimation method, the lowest RMSE value (0.13) was obtained when the test terminated at 60 items. When the test terminated at 20 items and with the standard error termination rule, RMSE values (0.21) were quite close to each other. Generally, RMSE values were slightly higher in the MLE estimation method compared to the EAP method. In conditions using the EAP ability estimation method, it was observed that RMSE values slightly decreased in larger samples when content balancing was not performed, while values were very close to each other when content balancing was performed. In conditions using the MLE ability estimation method, RMSE values were quite close to each other in small and large samples, whether content balancing was performed or not, and regardless of whether content areas were equally or differentially weighted (Table 2).

Correlations (r) between true and estimated ability levels were examined separately for two different sample sizes, three different termination rules, and content ratios used in content balancing, using different ability estimation methods. Accordingly, the highest correlation (r=0.99) between true and estimated ability levels for both sample sizes was obtained when the test terminated at 60 items, using both the EAP and MLE ability estimation methods. The fidelity coefficients obtained when the test was terminated at 20 items and with the standard error termination rule (SE≤.30) were quite close to each other. Content balancing did not affect the fidelity coefficients. It was observed that fidelity coefficients were slightly lower in conditions with differentially weighted contents compared to equally weighted content balancing (Table 2).

The effectiveness of whether content balancing was performed or not was also compared in terms of average number of items used in two different ability estimation methods. When the standard error termination rule (SE≤.30) was applied, the lowest average number of items (17.32) was obtained in conditions where the EAP ability estimation was used and content balancing was not performed. The average number of items was quite close across different sample sizes. The highest average number of

* Measurement and Evaluation Specialist, Ministry of National Education, Ankara, Türkiye, ilkayocal83@gmail.com, ORCID ID: 0009-0004-2246-6909
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Üçgül Öcal, İ., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications, *Journal of Measurement and Evaluation in Education and Psychology*, *15*(4), 394-407. https://doi.org/10.21031/epod.1438977

Received: 18.02.2024
Accepted: 18.12.2024

items was (19.75) in conditions where the MLE ability estimation was used and content balancing was performed with differentially weighted content areas. The average number of items was quite close across different sample sizes. Sample size did not affect the average number of items in either ability estimation method. Content balancing, in conditions with equally and differentially weighted content areas, increased the average number of items by approximately one item in conditions using both the EAP and MLE ability estimation methods.

The impact of content balancing on test security was also compared in terms of maximum item exposure (r_{max}) rates. The maximum item exposure rates obtained under 36 different conditions considered in the study are provided in Table 3.

Table 3

Maximum Item Exposure Rates (r_{max}) Under Different Conditions in Computerized Adaptive Testing Applications

Sample Size	Ability Estimation Method	Termination Rule	Content Balancing		
			None	Equally weighted	Differently weighted
250	EAP	20 items	0.5835	0.5593	0.6222
250	EAP	60 items	0.6767	0.6718	0.7006
250	EAP	SE≤0.30	0.5766	0.5428	0.5789
250	MLE	20 items	0.5580	0.5200	0.5521
250	MLE	60 items	0.6598	0.6526	0.6782
250	MLE	SE≤0.30	0.5522	0.5190	0.5382
500	EAP	20 items	0.5783	0.5349	0.6628
500	EAP	60 items	0.6704	0.6347	0.7311
500	EAP	SE≤0.30	0.5672	0.5225	0.6164
500	MLE	20 items	0.5370	0.5156	0.5727
500	MLE	60 items	0.6414	0.6326	0.6987
500	MLE	SE≤0.30	0.5334	0.5076	0.5696

In the small sample size, both the EAP and MLE ability estimation methods have shown that applying the termination at 20 items and the standard error termination rule (SE≤.30) reduced the maximum item exposure rate when content areas were equally weighted in content balancing. However, in the termination rule of stopping the test at 60 items, the rates are quite close to each other. In conditions where content balancing was done with differentially weighted content areas, the maximum item exposure rates increased with the EAP ability estimation method, whereas a decrease in this rate was observed when the MLE estimation method was used with the termination at 20 items and the standard error termination rule (SE≤.30).

In the large sample size, for both ability estimation methods, the maximum item exposure rates decreased in all conditions when content balancing was done with equally weighted content areas. In the case of content balancing with differentially weighted content areas, these rates increased in all conditions. Considering all conditions, the lowest item exposure rate (0.51) was observed in the large sample using the MLE estimation method with the standard error termination rule applied and when content areas were equally weighted in content balancing. The highest item exposure rate (0.73) was observed in the large sample using the EAP estimation method with the test termination rule at 60 items and when content balancing was done with differentially weighted content areas.

Discussion

Considering all findings obtained from the study, it has been observed that bias values, one of the indicators of measurement precision, slightly increase when content balancing involves differentially weighting content areas compared to other conditions. Generally, bias values were found to be lower in the EAP estimation method than in the MLE method. The RMSE value was not affected by whether content balancing was performed with equally or differentially weighted content areas when using the MLE estimation method. Without content balancing, RMSE values were quite close to each other in both small and large samples, regardless of whether content areas were equally or differentially weighted. Additionally, in conditions using the EAP estimation method, a slight decrease in RMSE values in larger samples was observed when no content balancing was performed. Generally, RMSE values were found to be slightly higher in the MLE method compared to the EAP method. An increase in the number of items reduced both RMSE and bias values, and the standard error termination rule and the termination at 20 items rules provided similar results. Regardless of sample size and ability estimation method, the highest correlation between true and estimated ability levels was obtained when the test terminated at 60 items. The selection of a 60-item test length in our study is supported by similar research and offers several advantages. Kingsbury et al. (2009) demonstrated that a 60-item exam allows for comprehensive content coverage and reliable, valid scores, equivalent to traditional tests of twice the length. Moreover, Sarı (2019) showed that longer tests mitigate adverse effects related to test security and reliability. Therefore, the 60-item length ensures adequate content representation and maintains high test reliability and validity, aligning with our study's goals. When content areas were differentially weighted, fidelity coefficients were found to be relatively lower compared to equal weighting. In both ability estimation methods, an increase in test length of about one item was observed when the standard error termination rule was applied. From this, it can be said that the increase in test length when content balancing is performed does not reduce test reliability to a significant extent. In all conditions, content balancing with equally weighted content areas reduced the maximum item exposure rates. Moreover, in the small sample, except for conditions where the test was terminated at 20 items and according to the SE<0.30 rule with the MLE method, content balancing with differentially weighted content areas increased the maximum item exposure rates. It can be said that content balancing conditions with equally weighted content areas perform better in terms of test security.

In synthesizing the outcomes of this study with those from related research, it's evident that the field of CAT is actively exploring the balance between measurement precision and content diversity. This study, alongside those by Leung et al.(2003b), Yasuda and Hull (2021), Yi and Chang (2010), and Zheng et al. (2013) collectively underscores the nuanced yet critical importance of content balancing in enhancing CAT's efficiency and accuracy without compromising item pool security and utilization. This study contributes to this body of knowledge by demonstrating that content balancing, while slightly increasing test length, does not detrimentally impact measurement precision. This finding aligns with Leung et al.'s (2003b) observation that certain item selection methods, notably the b-blocking method and MMM, optimize item pool utilization and minimize item overlap, suggesting that a thoughtful integration of stratification strategies and content balancing methods can achieve optimal outcomes in CAT applications. Moreover, the outcomes from Zheng et al. (2013) and Yasuda and Hull (2021) further reinforce the potential of content balancing strategies, such as the MMM, to effectively manage item exposure rates while maintaining test precision. This is particularly relevant in contexts requiring strict content specifications, where balancing can mitigate the risk of item overexposure without sacrificing measurement accuracy. Yi and Chang's (2010) introduction of a content-blocking method offers an innovative approach to item pool stratification, achieving balanced item usage and maintaining precision, which echoes this study's emphasis on the feasibility of content balancing in practical CAT designs. The collective findings suggest that while methodologies and focus areas may vary, the overarching goal remains consistent: refining CAT strategies to preserve the integrity of the testing process, optimize item pool usage, and ensure accurate and efficient measurement of abilities.

Comparing the outcomes of various studies on CAT, we observe diverse approaches and impacts of content balancing on measurement precision. Leung et al. (2003b) highlight how specific item selection methods like b-blocking method to multiple stratification and MMM optimize item pool utilization without affecting measurement accuracy, contrasting with our study's emphasis on the slight increase in

test length due to content balancing. Zheng et al. (2013) and Yasuda and Hull (2021) focus on content balancing's effect on specific domains or inventories, showcasing its variable impact on measurement precision. Yi and Chang's (2010) content-blocking method presents a novel approach, differing from traditional strategies by enhancing item pool usage efficiently. These differences underline the complexity of optimizing CAT, suggesting that the choice of content balancing strategy should be tailored to specific testing requirements and goals.

This study examined the effect of content balancing on measurement precision in dichotomous items under different measurement conditions in CAT applications. Control of item exposure rate, which holds significant importance in the CAT algorithm, was beyond the scope of this study. Future research could examine the impact of content balancing on measurement precision with control of item exposure rate. Similarly, the effect of content balancing when using different item selection methods could be explored. In the current study, the CCAT method was used as the content balancing method. Future studies could compare the performance of other content balancing methods on measurement precision using different packages or software.

Declarations

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

Author Contribution: İlkay ÜÇGÜL ÖCAL: conceptualization, investigation, methodology, data simulation, data analysis, supervision, writing - review & editing. Nuri DOĞAN: conceptualization, methodology, writing - original draft, formal analysis.

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 data has been simulated in this study.

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

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

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

References

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

Chen, S. K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, *58*(4), 569.

Chen, S.Y., & Ankenmann, R. D. (2004). Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing. Journal of Educational Measurement, *41*(2), 149–174. http://www.jstor.org/stable/1435211

Cheng, Y., & Chang, H-H. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467-482.

DeMars, C. (2010). *Item response theory*. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Demir, S. (2019). *Bireyselleştirilmiş bilgisayarlı sınıflama testlerinde sınıflama doğruluğunun incelenmesi*. [Doctoral Dissertation, Hacettepe Üniversitesi].

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah.NJ: Erlbaum.

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

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized Adaptive Testing: A primer* (2nd ed., pp. 37-60). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications*. Boston: Kluwer.

Han, K.T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*(7). https://doi.org/ 10.3352/jeehp.2018.15.7

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

Kalender, İ. (2009). Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: Bilgisayar ortamında bireyselleştirilmiş testler (Computerized adaptive tests-CAT). *CITO Egitim Kuram ve Uygulama*, *5*, 39-48.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359-375.

Kingsbury, G. G., Bontempo, B., & Zara A. R. (2009). *A comparison of CAT with LOFT methods for certification examinations.* [Conference presentation]. NOCA Annual Educational Conference.

Leung, C. K., Chang, H. H., & Hau, K. T. (2000). *Content balancing in stratified computerized adaptive testing designs* [Paper presentation]. AERA Annual Meeting, New Orleans.

Leung, C.K., Chang, H.H., & Hau, K.T. (2003a). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2*(5).

Leung, C.K., Chang, H.H., & Hau, K.T. (2003b). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*, *63*(2), 257-270.

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*(4), 287–304. https://doi.org/10.1207/s15324818ame0904_1

Özdemir, B., & Gelbal, S. (2015). İçerik ağırlıklandırmasının maddeler-arası boyutluluk modeline dayalı çok boyutlu bilgisayar ortamında bireyselleştirilmiş test yöntemleri üzerindeki etkisinin incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2). https://doi.org/10.21031/epod.03278

R Core Team (2013). *R: A language and environment for statistical computing*, (Version 3.0.1) [Computer software], Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters, In Weiss, D. (Ed). *New horizons in testing latent trait test theory and computerized adaptive testing*, 135-146. London: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50017-2

Rudner, L. M. (1998). An online, interactive, computer adaptive testing tutorial. Retrieved December 25, 2023, from https://edres.org/scripts/cat/catdemo.htm

Sari, H. İ., & Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice*, *17*(5). https://doi.org/10.12738/estp.2017.5.0484

Sarı. H. İ. (2019). Investigating consequences of using item pre-knowledge in computerized multistage testing. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi,* 39(*2*), 1113-1134. https://doi.org/10.17152/gefad.535376

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement*. New York: Academic Press.

Song, T. (2010). *The effect of fittıng a tridimensional irt model to multidimensional data in content-balanced computerized adaptive testing*. [Doctoral Dissertation, Michigan State University].

Şahin, A., Özbaşı D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive tests. *Eurasian Journal of Educational Research*, *17*(69), 21-36. http://dx.doi.org/10.14689/ejer.2017.69.2

Şenel, S. (2021). *Bilgisayar ortamında bireye uyarlanmış testler*. Pegem Akademi, Ankara.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*(5), 778-793. https://doi.org/10.1177/0013164408324460

Thompson, N. A., & Weiss, D. A. (2019). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1). https://doi.org/10.7275/wqzt-9427

Tian, J., Miao, D., Zhu, X., & Gong, J. (2007). An introduction to the computerized adaptive testing. *Us-China Education Review*, *4*(1), 72-81.

Van der Linden, W., & Glas, G. A. W. (2002). *Computerized adaptive testing: theory and practice*. Kluwer Academic Publishers.

Yasuda, J. I., & Hull, M. M. (2021). *Balancing content of computerized adaptive testing for the Force Concept Inventory* [Conference presentation]. Physics Education Research.

Yi, Q., & Chang, H-H. (2010). a-stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 359-378. https://doi.org/10.1348/000711003770480084

Zheng, Y., Chang, C-H., & Chang, H-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Qual Life Res*, *22*(3), 491-499. https://doi.org/ 10.1007/s11136-012-0179-6

Zheng, Y., & Chang, H-H. (2014). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*. *39*(2), 105-118. https://doi.org/10.1177/0146621614544519

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 408-420

The Effect of Missing Data Handling Methods on Differential Item Functioning with Testlet Data*

Rabia AKCAN ** Kübra ATALAY KABASAKAL ***

Abstract

This study examined the effect of three missing data handling methods (listwise deletion, zero imputation and fractional hot-deck imputation) on differential item functioning (DIF) with testlet data with a variety of sample size and missing data percentage under missing completely at random, missing at random, and missing not at random missing mechanisms. The study was conducted on two different datasets consisting of six testlets which contain 20 reading comprehension items of a foreign language test. Data with left-skewed distribution was referred to as data1 and data with right-skewed distribution was referred to as data2. In current study, false DIF was identified in data1 with all missing data methods under the missing at random mechanism with a 5% missing data rate in small sample size. Similarly, in analyses performed under the missing at random mechanism for data2, the proportion of items classified as false DIF was notably higher in the small sample size. Results also indicated that in all conditions, list wise deletion had the lowest correlations with DIF values obtained from the original datasets, datasets containing no missing data and serve as a reference for comparative analyses with datasets where missing data were artificially introduced. The zero imputation and fractional hot-deck imputation methods produced similar correlations when the missing data percentage was set at 5%. However, in the case of 15% missing data, zero imputation exhibited higher correlation values. Besides, in all conditions correlation values decreased with the increase of missing data percentage regardless of the missing data handling method.

Keywords:testlet, missing data, differential item functioning

Introduction

To date, there has been an ongoing investigation on defining and achieving validity. Validity can be defined as the degree to which evidence and theory support the test scores' interpretations for intended uses of tests. It is the most essential consideration for the development and evaluation of tests (AERA et al., 2014). Therefore, accumulating evidence for the validity of test scores is crucial for effective test development and evaluation.

Item bias is one of the key issues in test validity. It becomes evident when examinees of one group have a lower probability of success on the item than examinees of another group at the same ability level due to some characteristic of the test item or testing situation that is irrelevant to the test purpose (Zumbo, 1999). If any test item provides advantage to one of the groups, it negatively affects validity. Therefore, bias studies can play an important role in addressing the issue of validity.

The first step in identifying item bias is to detect items containing differential item functioning (DIF). DIF occurs when examinees from different groups, who have been matched on the ability of interest, have differing probabilities or likelihoods of succeeding on an item (Clauser & Mazor, 1998). DIF is

*This study was produced from the doctoral dissertation conducted by the first author under the supervision of the second author.

**Teacher., Republic of Turkey Ministry of National Education, Afyonkarahisar-Turkey, eltrabia42@hotmail.com, ORCID ID: 0000-0003-3025-774X
***Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Turkey, katalay@hacettepe.edu.tr, ORCID ID: 0000-0002-3580-5568

To cite this article:
Akcan, R., Atalay Kabasakal, K., (2024). The effect of missing data handling methods on differential item functioning with testlet data. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(4), 408-420. https://doi.org/10.21031/epod.1539940

Received: 28.08.2024
Accepted: 24.11.2024

required, but not sufficient condition for item bias. If DIF is present, follow-up item bias analyses (e.g., content analysis) would have to be done to determine whether item bias is evident or not (Zumbo, 1999). In other words, DIF is a statistical technique which helps identifying potentially biased items.

DIF analyses play a vital role in test development and validation to assure that scores obtained from educational tests and psychological measures are not biased and they measure the same construct for all examinees (Walker, 2011). Although standalone item DIF analysis has received considerable attention, small bundles of items are the fundamental building blocks of many exams. A bundle is defined as any group of items chosen in accordance with some organizing principle. These items do not have to be adjacent or it is not necessary for them to refer to a common passage or a text (Douglas et al., 1996). For example, three independent math items based on analytical reasoning can form an item bundle in a test.

DIF analyses can be carried out both at the item and bundle level. Item bundle DIF, which is called as differential bundle functioning (DBF), is an extended form of item DIF (Douglas et al., 1996). As previously stated, items in a bundle are not necessarily close to each other or they do not have to share a common passage. However, Beretvas and Walker (2012) point out that there are various reasons to put the items together in a bundle. One of them is the testlets in which items might be bundled together. A testlet is a set of items which are based on a common stimulus (Wainer & Kiely, 1987). For instance, items within a testlet may focus on a laboratory scenario, a graphic, a reading passage or complex problem (DeMars, 2006). Testlets save testing time since examinees focus on the scenario once and they can utilize the information for other items. Besides, authenticity of the task may increase as more context is added (DeMars, 2012). A well-known example of testlets is reading comprehension items which are based on a paragraph in language tests. The difference between a testlet and an item bundle is that items in a testlet share a common input whereas this is not the case for an item bundle. Therefore, examining item bias in testlets with a different method provides more valid and reliable results.

SIBTEST (Shealy & Stout, 1993) and Poly-SIBTEST (Chang et al., 1996), which is an extended form of SIBTEST for polytomous items, have been commonly used in the detection of DIF at the item level and DBF at the testlet level (Beretvas & Walker, 2012; Lee, Cohen & Toro, 2009; Min & He, 2020). DIF can only be identified at the testlet level when DBF analysis is carried out within the framework of SIBTEST method, proposed by Douglas et al. (1996), and thus may be referred as differential testlet functioning. In this case, it is not possible to determine which items are causing differential testlet functioning. Moreover, creating a testlet is expensive and time-consuming. It's better to handle problems at the item level by determining problematic items instead of discarding the whole testlet from the item bank due to differential testlet functioning. Accordingly, it would be more practical and useful to apply a method which examines DIF at the item level rather than a method examining differential testlet functioning (Fukuhara & Kamata, 2011).

A Bifactor MIRT Model ForTestletsWith Covariates

Fukuhara and Kamata (2011) proposed a DIF detection model which is an extension of a bifactor multidimensional item response theory (MIRT) model for testlets. Unlike conventional item response theory (IRT) DIF models, this proposed model takes testlet effects into consideration. Consequently, it estimates DIF magnitude appropriately if a test consists of testlets. Moreover, DIF can be identified for all items simultaneously with the proposed DIF model. It also estimates DIF magnitudes assuming that the average DIF magnitude is zero. Besides, there is a parameter to capture the mean ability difference between the focal and reference groups to distinguish DIF and impact. If this parameter is not included, it is assumed that no ability difference between the focal and reference groups exists. The bifactor MIRT model for testlets with covariates proposed by Fukuhara and Kamata (2011) is reduced to a traditional IRT model if there is no testlet effect, which ensures that there will be no adverse impact when the testlet effect is not present. The authors set the absolute value of 0.426 in the logit scale as the threshold for a meaningful DIF magnitude in their study. The current study utilized

the proposed model by Fukuhara and Kamata (2011) for the DIF detection process and the absolute value of 0.426 was adopted to identify meaningfully large DIF items as the researchers did in their study.

Missing Data

Another significant aspect of validity is missing data which may cause incorrect trait inferences, thereby reducing validity (Garrett, 2009). Any blank responses to the entire set of items that a test taker has access to are referred to as missing data (Ludlow & O'leary, 1999). In real life assessment scenarios, missing data are widely seen. There are various methods to handle missing data. Type of missing data and the method chosen to handle missing data can have a unique impact on the statistical results (Garrett, 2009).

In order to decide on the appropriate method to handle missing data, one must address missing data type. Rubin (1976) classified missing data mechanisms into three types: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Data are MCAR when there is no relation between the probability of missing data on a variable Y and the value of Y itself or the values of any other variables in the dataset (Allison, 2002). MAR data is present if the probability of missing data on the variable Y is related to any other variables in the model but it is not related to the values of Y itself (Enders, 2010). In other words, data are MAR when the probability of missing data is just influenced by the values of other observed variables (Robitzsch & Rupp, 2009). Data are MNAR if the probability of missingness on the variable Y is related to the values of Y itself, even after controlling for other observed variables (Enders, 2010). Missingness cannot be explained with observed variables for MNAR data, as it depends on the unobserved values (Robitzsch & Rupp, 2009).

In real life situations DIF and missing data might occur at the same time. It is essential to investigate DIF in the presence of missing data. Nevertheless, commonly used DIF detection methods such as Mantel Haenszel (MH), SIBTEST and Logistic Regression (LR) cannot handle missing data (Banks, 2015). To be more specific, In MH DIF analysis for example, students' abilities are typically matched based on the number of items they answer correctly. Since this matching is based on the number correct answers, missing items are generally considered as either not administered or incorrectly answered. Thus, it is crucial to investigate how these methods and other missing data handling methods affect DIF analysis (Emenogu et al., 2010).

Missing data handling methods applied for the analysis may also lead to bias. Choice of missing data method may create DIF when there is no DIF in the item or eliminate DIF when it is actually present (Banks, 2015). There have been research on DIF detection in the presence of missing data with simulated data (Finch, 2011a; Finch, 2011b; Garrett, 2009; Robitzsch & Rupp, 2009) or real data (Rousseau et al., 2004; Tamcı, 2018). Most of these studies have focused on DIF and missing data in different aspects. However, very little attention has been paid to the role of missing data on DIF detection with testlet data.

Since items in a testlet are dependent, when an examinee leaves an item blank in the testlet, responses to other items will probably be affected, which creates MNAR data. If the examinees from one group leave the items blank at a larger rate than those from the other group, this produces unbalanced data (Sedivy, 2009). As a result, it is important to investigate the impact of missing data and missing data handling methods on DIF detection with testlet data.

A number of techniques have been developed to handle missing data problem. The present study used listwise deletion (LD), zero imputation (ZI) and fractional hot-deck imputation (FHDI) to deal with missing data. In LD, any observations with missing data on the variables are deleted from the sample. One advantage of LD is that it can be applied for any type of statistical analysis. Another advantage of it is that it does not require any special computational technique (Allison, 2002). However, there are certain drawbacks associated with the use of LD. Primarily, it requires MCAR data and may create inaccurate parameter estimates when this assumption is ignored. It can also produce biased estimates when the data are only MAR, but not MCAR. Aside from bias, if discarded observations have data on

many variables, discarding observations with missing data reduces total sample size drastically and thus causes decrease in statistical power (Allison, 2002; Enders, 2010). In literature, however, LD has been widely used in research investigating missing data and DIF together (Banks & Walker, 2006; Emenogu et al., 2010; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Sedivy et al., 2006). As already stated, in real life situations dependency of the items in a testlet is likely to cause MNAR data if an examinee leaves an item blank, so it would be interesting to see how LD method works under three different missing data mechanisms on DIF detection with testlet data.

ZI is one of the most basic techniques for imputing item response data among the techniques that employ a single imputation step. In this method, all missing values are replaced with a score of zero. Nonetheless, some researchers do not consider this method as a true imputation method because it lacks a statistical model. However, it is frequently used in the context of achievement tests because it is easy to do and can reasonably suggest that a lack of response shows lack of proficiency (Robitzsch & Rupp, 2009).

FHDI, proposed by Kalton and Kish (1984) and investigated by Kim and Fuller (2004), is a way of performing hot deck imputation efficiently. In this method, M imputed values are produced for each missing value as in multiple imputation (MI); nevertheless, a single dataset is obtained as the output after fractional imputation (Im et al., 2015). The imputed values are randomly selected from the donors' data within the same imputation cell. These cells are particularly created to ensure data homogeneity within each cell. Fractional weights are assigned to each imputed value to maintain the original data structure and for variance estimation replication methods are adopted (Im et al., 2015; lm et al., 2018). FHDI was extended by Im et al. (2015) in two ways. First, in this new version of FHDI, a nonparametric imputation approach, imputation cells are not required to be made in advance. Instead, multiple cells are allowed for each missing item. Second, the proposed FHDI method is applied to multivariate missing data with arbitrary missing patterns. In the current study, we utilized extension of FHDI proposed by Im et al. (2015) as the imputation method. Fractional imputation has not yet seen widespread adoption outside of survey sampling, probably because it is a relatively new method and involves more complex variance estimation procedures compared to MI (lm et al., 2018). However, fractional imputation provides consistent variance estimates, especially when using a method-of-moment estimator. In contrast, MI may sometimes yield inconsistent variance estimates (Yang and Kim, 2016). FHDI approach used in this study utilizes observed values as imputed values, and thus can better preserve the structure of the data. Despite the widespread use of MI, only one of the two methods was employed in this study due to practical constraints-specifically the extended analysis time required for DIF analyses. FHDI was selected for this study due to its noted advantages and its potential as a promising method.

Research on DIF in the presence of missing data has produced various results: In a simulation study Finch (2011a) found that compared with LD and MI, ZI had highly inflated type I error rates under MAR mechanism and it was found to be the least applicable method under this condition. Results also indicated that LD and MI performed similarly. Another simulation study by Finch (2011b) also demonstrated that LD was superior to ZI under various conditions. In their simulation study, Robitzsch and Rupp (2009) concluded that incorrect choice of missing data method led to false DIF. In addition, missing data handling methods had less problematic results under MCAR mechanism. Several studies used real data to investigate the impact of missing data handling methods on DIF detection. Akcan and Atalay Kabasakal (2023) focused on the impact of missing data on DIF detection using LD, ZI and FHDI methods under MCAR mechanism. They reported that FHDI yielded the best results in detecting DIF items in all conditions and DIF values obtained with FHDI were the closest DIF values to those obtained from the original dataset. Tamcı's (2018) study on DIF detection with a variety of DIF magnitude, sample size and focal/reference group rate in case of MCAR data showed that MI had lower error rates than ZI and expectation-maximization. Power rates for these methods were mostly below the acceptable level with the exception of ZI for 10% missing data percentage. Emenogu et al. (2010) investigated the impact of ZI, LD and analysis wise deletion on MH method with both real and simulated data. They reported that ZI produced false DIF regardless of the matching criterion used in the study and LD led to a significant decrease in sample size and the power of MH method.

Nichols et al. (2022) assessed DIF on a real dataset by using single hot-deck imputation and Multiple Imputation by Chained Equations (MICE) as a multiple imputation technique with a variety of missing rate and DIF scenarios. They reported that MICE achieved slightly better results than hot deck single imputation in reducing observed DIF estimation errors, although both methods were effective in decreasing observed errors compared to scenarios without any imputation. They suggested using MICE in testing DIF to reduce the bias caused by missing data when the missing data rate exceeds the 10% threshold. They also stated that MICE could not remove the observed error due to missing data in their study. As a result, they advised investigators to interpret results with caution when they employ MICE to handle missing cognitive data.

Purpose of the Study

Testlets are widely utilized in many high-stakes testing situations (e.g. American College Testing, Graduate Record Examination, Test of English as a Foreign Language and International English Language Testing System). Various studies have investigated DIF in testlet-based items to determine the effect of testlets on DIF detection (Fukuhara & Kamata, 2011; Min & He, 2020; Ravand, 2015; Sedivy, 2009; Taşdelen Teker, 2014; Wang & Wilson, 2005). It is inevitable that there will be missing data in real life situations. Although testlets are widely used in large scale examinations, there have been no attempts to investigate the effect of missing data handling methods on DIF with testlet data. Determining the conditions that missing data handling methods work best will contribute to the accuracy of DIF detection results. Therefore, this study provides new insights into DIF detection with testlets in the presence of missing data. The leading research question in this investigation is as follows: Do missing data handling methods have an impact on DIF detection with testlet data with a variety of sample size and missing data percentage? To achieve the goal of this research, following sub-problems are addressed:

1) How do DIF results change across sample size (1,000 and 2,000) and missing data percentage (5% and 15%) under MCAR, MAR and MNAR mechanisms by using LD, ZI and FHDI missing data handling methods?

2) How do the correlations between DIF magnitudes obtained from the original datasets and new datasets change across sample size (1,000 and 2,000) and missing data percentage (5% and 15%) under MCAR, MAR and MNAR mechanisms by using LD, ZI and FHDI missing data handling methods?

Method

Dataset

The dataset used in this research was obtained from students' responds to six testlets composed of 20 items in English test of Undergraduate Placement Exam (UPE) conducted in Turkey in 2016. Items that formed testlets were cloze test (Items 1-5 where the participants have to fill in the blanks in one reading passage), reading-1 (Items 6-8), reading-2 (Items 9-11), reading-3 (Items 12-14), reading-4 (Items 15-17) and reading-5 (Items 18-20). To get a complete dataset composed of testlets, all examinees having missing values were deleted from testlet data and a dataset consisted of 33570 examinees was created. Four schools out of 87 were chosen as the sample because they had sufficient sample size for the purpose of this study. These schools were private high schools teaching in foreign language (4275 students), science high schools (1183 students), religious high schools (2333 students) and formal high schools (4891).Two different datasets were created from these schools according to their distributions. Data distribution was regarded as another condition. Data1 consisted of private high schools teaching in foreign language and science high schools which had left-skewed distributions. Data2, on the other hand, consisted of religious high schools and formal high schools which had right-skewed distributions. Two sample size conditions (1,000 and 2,000) were included in

the study which accounted for four samples in total. These four samples were referred as original datasets.

Data Analysis

Datasets used in the study were data1 and data2, which had left-skewed and right-skewed distributions, respectively. Study was conducted on four samples (1,000 and 2,000 sample size for each) which were drawn from these datasets. To begin the process, DIF analyses were performed on four samples by using the bifactor MIRT model for testlets with covariates and results were used as reference. Missing data were generated under MCAR, MAR and MNAR mechanisms and percentage of missing data was set at 5% and 15%, and thus 24 datasets including missing data were created from the four original samples. Following this missing data generation process, LD, ZI and FHDI were adopted to handle missing data problem of 24 datasets and 72 datasets without missing data problem were obtained. Finally, DIF analyses were conducted on 72 datasets by using the bifactor MIRT model for testlets with covariates to compare the results with those obtained from the original datasets in the first stage of the process. LD and ZI were performed by writing codes on base R and "FHDI" (lm et al., 2018) package was used for imputation with FHDI. Testlet DIF analyses for all datasets were carried out using WinBUGS 1.4.3 (Spiegelhalter et al., 2003).

Missing Data Generation Process

Missing data were generated on the items under MCAR, MAR and MNAR mechanisms. Percentage of missing data was set at 5% and 15% for the entire dataset. In case of MCAR data, appropriate proportion of responses on all items from both reference and focal groups were randomly selected. For MAR data, responses on all items were randomly selected only from the focal group. Percentage of missing data deleted from the focal group in MAR case was set at 5% and 15% of the entire dataset. Under MNAR mechanism, missing data were generated in both groups and it depended on the item difficulties and total scores. Yet, percentage of missing data was set at the same proportion as MCAR and MAR data for the entire dataset. In MNAR missing data generation process, total scores were divided into three levels from lowest to the highest whereas item difficulties were divided into three levels from easiest items to the most difficult ones. Certain percentages of data were deleted in each level, total amount of which was equal to 5% or 15% of the entire dataset. To clarify, the amount of missing data was greater for the examinees with the lowest ability levels on the most difficult items compared to the examinees with the highest ability levels and so on. Missing data were created in R software by adapting the codes written by Doğanay Erdoğan (2012) to this study.

DIF Analyses

DIF analyses for all datasets were carried out using the bifactor MIRT model for testlets with covariates. Parameters were estimated using WinBUGS 1.4 program with a Markov chain Monte Carlo (MCMC) method. When the MCMC method is used to estimate parameters, it should be checked whether the parameter estimates converge. If convergence of the parameters is not achieved, incorrect inferences regarding the parameter of interest will be drawn. Therefore, it is necessary to decide the number of iterations to remove (i.e., burn-in iterations) when a parameter estimation becomes stable. Besides, the numbers of iterations after a burn-in period needs to be decided to get good samples of each parameter that represent the parameter's posterior distribution (Fukuhara & Kamata, 2011). In this study, preliminary analyses were conducted on the original datasets, and the burn-in iterations and total number of iterations were determined to achieve convergence. Based on the preliminary analysis, 9,800 samples were drawn from each posterior distribution after discarding 200 samples as burn-in period. Fukuhara and Kamata (2011) assessed convergence using history plots, density plots and auto-correlation plots in their study. Another way of assessing convergence is to check MC errors. Small values of MC errors show that parameter of interest is estimated accurately

(Ntzoufras, 2009). To assess convergence the present study used graphical methods (history, density and autocorrelation plots) and also checked MC errors. Results indicated that the parameter estimates of DIF converged.

Results

The first set of analyses determined the DIF items in original datasets. The bifactor MIRT model for testlets with covariates proposed by Fukuhara and Kamata (2011) was used to identify DIF and the value of 0.426 was adopted to identify meaningfully large DIF items as the researchers did in their study. Table 1 presents DIF items in each original sample.

Table 1
DIF results in original datasets.

	Data1		Data2	
Item No	**1000**	**2000**	**1000**	**2000**
1	0.080	0.016	0.393	**0.455***
2	0.044	-0.115	-0.178	-0.084
3	**-0.474***	**-0.484***	-0.326	-0.260
4	0.164	0.006	**0.439***	**0.480***
5	0.071	0.025	-0.051	-0.079
6	-0.028	-0.052	-0.336	-0.105
7	-0.149	-0.105	-0.379	-0.175
8	-0.032	-0.007	-0.246	-0.184
9	-0.080	0.058	0.270	0.048
10	-0.129	-0.113	0.051	0.093
11	-0.184	-0.081	-0.161	-0.106
12	-0.024	-0.076	**-0.526***	-0.217
13	0.174	0.120	0.290	0.198
14	0.067	0.125	-0.059	-0.113
15	0.049	0.052	0.148	0.073
16	0.382	0.235	0.424	**0.435***
17	-0.056	0.047	-0.078	-0.172
18	0.194	0.173	0.209	-0.201
19	-0.153	0.001	-0.099	-0.171
20	0.084	0.177	0.214	0.085

**DIF magnitude>0.426*

It is apparent from Table 1 that only item 3 showed DIF in both sample sizes in data1. The other four items (items 1, 2, 4 and 5) in the same testlet with item 3 did not display significant DIF and their magnitudes were quite low. Results obtained from data2 indicated that only one DIF item (item 4) was common in the two sample sizes. Two items (items 4 and 12) were flagged DIF in small sample whereas three items (items 1, 4 and 16) showed DIF in the larger sample. In addition, some non-DIF items (items 1, 7 and 16) in small sample size had DIF magnitude relatively close to the value of 0.426.

Table 2 provides DIF items in all conditions for data1 after the treatment with ZI, LD and FHDI methods. At the beginning there were 36 conditions in total, however, eight of them could not be carried out because of the reduced sample size with LD method and presence of missing data in all cases for the focal group with FHDI method.

Table 2
DIF items in all conditions for data1 with ZI, LD and FHDI methods.

Method	Item No.					
	5%			15%		
	MCAR	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
1000						
ZI	3	3,16	3	3	1, 3, 4,16	3
LD	*	16	3	-	-	-
FHDI	3	3,16	3	3,16	-	3, 5,18
2000						
ZI	3	3	3	3	3, 4,16	3
LD	3	3,20	*	-	-	-
FHDI	3	3	3	3,18	-	3, 10,16

**None of the items displayed DIF.*

What stands out in Table 2 is ZI and FHDI methods were superior to LD in detecting DIF item (item3) when the percent of missing data was set 5%. It was also found that the effects of three missing data handling methods on the performance of identifying DIF free items were similar. Another important finding was that under MAR mechanism in small sample size case, item16 showed false DIF with the three missing data handling methods.

For 15% missing case, only ZI and FHDI results were compared. As can be seen from the table, item3 was correctly identified as DIF item in all conditions. However, under MCAR and MNAR mechanisms, ZI performed better than FHDI method in terms of identifying DIF free items in the original datasets. On the other hand, ZI produced false DIF under MAR mechanism and the percentage of items that showed false DIF with ZI was found to be higher in smaller sample size. Closer inspection of the table shows that two items (item1 and item3) displaying false DIF in small sample under MAR mechanism with ZI in 15% missing case were in the same testlet. Likewise, two items (item3 and item5) displaying false DIF in small sample under MNAR mechanism with FHDI in 15% missing case were in the same testlet.

DIF items in all conditions for data2 after the treatment with ZI, LD and FHDI methods are presented in Table 3. At the beginning there were 36 conditions in total, however, six of them could not be carried out because of the reduced sample size with LD method and presence of missing data in all cases for the focal group with FHDI method.

Table 3
DIF items in all conditions for data2 with ZI, LD and FHDI methods.

Method	Item No.					
	5%			15%		
	MCAR	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
1000						
ZI	4,12	7, 12,13	4,12	1, 4, 7, 12,16	6, 7, 10, 12, 13,16	*
LD	1, 2, 7, 12, 13,18	8,12	*	-	-	8, 15, 20
FHDI	4,12	1,12	12	1, 4, 6, 8,12, 13, 16, 18, 19,20	-	*
2000						
ZI	1,4	*	1,4	1, 4,16	4, 10, 13,16	1, 4
LD	4,20	*	*	-	-	4, 10
FHDI	1,4	4	4,16	1, 2, 4,16	-	4

**None of the items displayed DIF.*

From this data presented in Table 3, we can see that the percentage of items which had false DIF in small sample size was greater than or equal to those from the large sample size regardless of the missing data rate. Particularly, in analyses performed under the MAR mechanism, the proportion of items classified as false DIF is notably higher in the small sample size. As regards to the detection of DIF items in the original datasets, the results demonstrated that performances of ZI and FHDI were similar. Yet, these two methods led to false DIF in some conditions for both missing data percentage. Of interest here is the increase in error rate of classifying DIF and non-DIF items when the percentage of missing data was set 15%. Results also indicated that when the percentage of missing data was set 5% in case of MCAR and MNAR data, LD had the lowest performance on detecting DIF items regardless of the sample size. On the other hand, performance of identifying DIF free items in eight conditions improved with the increase in sample size.

It is apparent from Table 3 that some of the items that displayed false DIF (e.g. 18-20) were within the same testlets. It was also found that with ZI method, items that are in the same testlet and displayed false DIF together were mostly observed in case of MAR data. Table 4 shows the Pearson correlations between the DIF magnitudes obtained from all conditions and the ones obtained from the original datasets.

Table 4

Correlations between the DIF magnitudes obtained from all conditions and the original datasets.

Method		**5%**			**15%**	
	MCAR	**MAR**	**MNAR**	**MCAR**	**MAR**	**MNAR**
Data1_1000						
ZI	0.92**	0.91**	0.97**	0.89**	0.80**	0.85**
LD	0.83**	0.74**	0.60**	-	-	-
FHDI	0.98**	0.99**	0.97**	0.82**	-	0.72**
Data1_2000						
ZI	0.98**	0.90**	0.98**	0.95**	0.80**	0.94**
LD	0.88**	0.83**	0.73**	-	-	-
FHDI	0.97**	0.98**	0.89**	0.74**	-	0.77**
Data2_1000						
ZI	0.99**	0.97**	0.99**	0.97**	0.84**	0.93**
LD	0.88**	0.52*	0.90**	-	-	0.20
FHDI	0.98**	0.97**	0.99**	0.73**	-	0.92**
Data2_2000						
ZI	0.99**	0.95**	0.99**	0.97**	0.76**	0.96**
LD	0.71**	0.83**	0.64**	-	-	0.58**
FHDI	0.98**	0.98**	0.99**	0.68**	-	0.96**

**p<.05; **p<.01.*

From this data, we can see that LD method resulted in the lowest correlations in all conditions for both data1 and data2. Furthermore, in all conditions for both datasets there was a decrease in correlation values as the percentage of missing data increased. ZI and FHDI had similar results when the missing data percentage was set 5%. However, for 15% missing case ZI method had higher correlation values. Lowest correlation values with ZI were obtained under MAR mechanism.

The correlation results for data1 showed that when the sample size increased, a higher correlation was obtained with the LD method. Yet, a similar pattern was not obtained from the results of data2. It was determined that except for one condition, the correlation values obtained from data2 with the ZI method were higher than the values obtained from data1, which was not the case for LD and FHDI methods. The reason for this might be that data2 was skewed to the right, and thus DIF magnitudes were less affected by the imputation with ZI method.

Discussion and Conclusion

This study set out to investigate the impact of missing data handling methods on DIF detection with testlet data with a variety of sample size and missing data percentage. Study was conducted on four samples (1,000 and 2,000 sample size for each) which were drawn from two different datasets. For LD method, discarding observations with 15% missing data led to a significantly reduced sample size in most of the conditions, and thus DIF parameters could not be estimated with LD method under this condition. The study was limited to the results obtained with lower percentage of missing data with LD method. This finding is consistent with that of Emenogu et al. (2010).

Results obtained from both datasets showed that LD method produced the lowest correlations with reference DIF values in all conditions. While LD method performed as efficiently as or slightly lower than ZI and FHDI in detecting DIF free items, ZI and FHDI were superior to LD in detecting DIF items in many conditions for both datasets. This finding seems to be consistent with the research by Akcan and Atalay Kabasakal (2023).

Results from data1 indicated that in all conditions, there was an increase in error in detecting DIF free items with FHDI method as the percentage of missing data increased. It was also revealed that performance of ZI method in detecting DIF free items under MAR mechanism was adversely affected by the increase in missing data percentage. Likewise, there were conditions in data2, in which performance of detecting DIF free items decreased with the larger missing data percentage in three missing data mechanisms. In addition, correlation values obtained from the three missing data handling methods in all conditions decreased, as the percentage of missing data increased. These results are in agreement with those obtained by Emenogu et al. (2010). They stated in their research that impact of missing data handling method was insignificant when the missing data percentage was low. However, percentage of missing data in focal or reference group might be a source of DIF when the percentage of missing data was large.

Finch (2011a) reported that type I error rate of ZI method was greatly inflated in all conditions under MAR mechanism, which was not the case for LD and MI methods. The present study found that performance of identifying DIF free items under MAR mechanism was lower than MCAR and MNAR mechanisms in all conditions except for the results obtained from 2,000 sample size from data1 with ZI and FHDI methods. The two researches had similarities in this respect.

According to the correlations with reference DIF values, ZI and FHDI produced similar results in both datasets when the missing data percentage was set at 5%. On the other hand, correlations with ZI were higher than FHDI in case of 15% missing data. In their research, Akcan and Atalay Kabasakal (2023) reported that FHDI had the highest correlation values in all conditions and ZI had slightly lower correlation values than FHDI. This differs from the results presented here. A possible explanation for this might be the different distributions of the data used in these studies. Especially right skewed distributed datasets are likely to be less affected by imputation with ZI and produce closer DIF values to the values obtained from the original datasets.

Another finding was that performance of identifying DIF free items in eight conditions in data2 improved with the increase in sample size. This study also found that under MAR mechanism, there was a decrease in error in detecting DIF free items with ZI and FHDI methods when the sample size increased. These results are in agreement with Tamcı (2018) who showed that ZI yielded unacceptable type I error rates when the sample size decreased.

This current study was limited by two sample size as datasets used here were drawn from a real dataset. Missing data percentage was set at 5% and 15% due to the relatively small sample sizes. Nevertheless, it was not possible to estimate DIF parameters in some conditions with 15% missing data case because of the reduced sample size with LD method and presence of missing data in all cases for the focal group with FHDI method. The study did not include any other DIF detection methods for testlets. In future research, it would be beneficial to employ multiple DIF detection methods so that researchers can gain insights into how various DIF detection methods work in presence of missing data. Further research might also explore the impact of different missing data handling methods on

testlet DIF with a larger sample size and missing data percentage. The study can be repeated using both real data and simulated data using different DIF detection methods and missing data handling methods. A further study might also investigate DIF in the presence of missing data with testlets and the standalone items together.

Declarations

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

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

AERA, APA, and NCME (2014). *The standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Akcan, R., & Atalay Kabasakal, K. (2023). The impact of missing data on the performances of DIF detection methods. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 95-105. https://doi.org/10.21031/epod.1183617

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA, US: Sage publications.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, *20*(12), 1-10.

Banks, K., & Walker, C. (2006). *Performance of SIBTEST when focal group examinees have missing data*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement*, *72*(2), 200-223. https://doi.org/10.1177/0013164411412768

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *ETS Research Report Series*, *1995*(1), i:30.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement Issues and Practice*, *17*(1), 31-44. https://eric.ed.gov/?id=EJ564712

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104-121. https://doi.org/10.1177/0146621612437403

Doğanay Erdoğan, B. (2012). *Assessing the performance of multiple imputation techniques for Rasch models with a simulation study*[Doctoral dissertation, Ankara University]. Council of Higher EducationThesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=RjkkLNSzxvPF7_el9Z6dkg&no=6IGSpCmQZHcSkxQALa295w

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, *33*(4), 465-484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Emenogu, B. C., Falenchuk, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research*, *56*(4), 459-469. https://doi.org/10.11575/ajer.v56i4.55429

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

Finch, H. (2011a). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education*, *24*, 281-301. https://doi.org/10.1080/08957347.2011.607054

Finch, H. (2011b). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, *71*(4), 663-683. https://doi.org/10.1177/0013164410385226

Fukuhara, H., &Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622. https://doi.org/10.1177/0146621611428447

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size*.) [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/eps_diss/35/

Im, J., Cho, I. H., & Kim, J. K. (2018). *FHDI: Fractional Hot Deck and Fully Efficient Fractional Imputation*. https://cran.r-project.org/web/packages/FHDI/index.html

Im, J., Kim, J. K., & Fuller, W. A. (2015). Two-phase sampling approach to fractional hot deck imputation. *In Proceedings of the Survey Research Methods Section*, 1030-1043.

Kalton, G., & Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, *13*(16), 1919-1939. https://doi.org/10.1080/03610928408828805

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, *91*(3), 559-578. https://doi.org/10.1093/biomet/91.3.559

Lee, Y.-S., Cohen, A., & Toro, M. (2009). Examining type I error and power for detection of differential item and testlet functioning. *Asia Pacific Education Review*, *10*(3), 365-375. https://link.springer.com/article/10.1007/s12564-009-9039-7

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational and Psychological Measurement*, *59*(4), 615-630. https://doi.org/10.1177/0013164499594004

Min, S., & He, L. (2020). Test fairness: Examining differential functioning of the reading comprehension section of the GSEEE in China. *Studies in Educational Evaluation*, *64*. https://doi.org/10.1016/j.stueduc.2019.100811

Nichols, E., Deal, J. A., Swenor, B. K., Abraham, A. G., Armstrong, N. M., Bandeen-Roche, K., Carlson, M.C., Grisworld, M., Lin, F. R., Mosley, T. H., Ramulu, P. Y., Reed, N. S., Sharrett, A. R., & Gross, A. L. (2022). The effect of missing data and imputation on the detection of bias in cognitive testing using differential item functioning methods. *BMC Medical Research Methodology*, *22(*1), 1-12. https://doi.org/10.1186/s12874-022-01572-2

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. John Wiley & Sons.

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015585607

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, *69*(1), 18-34. https://doi.org/10.1177/0013164408318756

Rousseau, M., Bertrand, R., & Boiteau, N. (2004). *Impact of missing data on robustness of DIF IRT-based and non IRT-based methods.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 2004.

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

Sedivy, S. K. (2009). Using traditional methods to detect differential item functioning in testlet data. [Doctoral dissertation, University of Wisconsin-Milwaukee]. ProQuest Dissertations Publishing. https://www.proquest.com/openview/4ea81f321746d15a968b1505d7c8102b/1?pq-origsite=gscholar&cbl=18750

Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). *Detection of differential item functioning with polytomous items in the presence of missing data*. Paper presented at the annual meeting of the National Council of Measurement in Education.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/ DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*,*58*(2), 159-194. https://doi.org/10.1007/BF02294572

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS version 1.4.3 [Computer Program]. MRC Biostatistics Unit, Institute of Public Health.

Tamcı, P. (2018). *Investigation of the impact of techniques of handling missing data on differential item functioning*. [Master's Thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=TuYJvxte3q-jTBLlv22wLg&no=hOhq2SUN1BT96zGOfFKULw

Taşdelen Teker, G. (2014). *The effect of testlets on reliability and differential item functioning*. [Doctoral Dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201. https://www.jstor.org/stable/1434630

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment,29*(4), 364-376. https://doi.org/10.1177/0734282911406666

Wang, W.-c., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*(4), 549-576. https://doi.org/10.1177/0013164404268677

Yang, S., & Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika, 103*(1), 244-251. https://doi.org/10.1093/biomet/asv073

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 421-445

Natural Language Processing and Machine Learning Applications For Assessment and Evaluation in Education: Opportunities and New Approaches

Kübra YILMAZ* Kaan Zülfikar DENİZ**

Abstract

This study examines the applications of Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP) technologies in education, particularly in educational assessment and evaluation processes. The study examines the potential of these technologies to contribute to educational assessment and evaluation processes in areas such as automatic item generation, text mining, sentiment analysis, sentence similarity, and providing feedback to students. The study includes both a literature review and sample applications. In the automatic item generation process of the study, language models such as GPT and Gemini are used to generate new educational questions and this process is supported by NLP technologies. The study is enriched with Turkish examples and the results show that these applications can be further developed for Turkish and have potential for other applications.

Keywords: machine learning in education, natural language processing in education, artificial intelligence, educational technologies

Introduction

AI technologies, which continue to develop today, have started to be used in many areas of life. Studies are being conducted on the integration of AI technologies into disciplines such as healthcare, law, architecture, and education, as well as on preventing various risks (Chaudhry & Kazim, 2021; Shin, 2021; Başarır, 2022; Lin, 2023; Ramachandran & Rana, 2024; İlikhan et al.,2024). AI is defined as "systems that perform given complex tasks by imitating human problem-solving abilities" (Newell & Simon, 1956). It is the science and engineering of creating intelligent machines (McCarthy, 2007). In this context, AI is often equated with algorithms. However, the term algorithm is a concept that existed before AI. The term algorithm is derived from the name of the Persian mathematician Muhammad ibn Musa al-Khwarizmi in the 9th century and means instructions developed to perform a calculation or solve a problem (Sheikh et al.,2023).

Russel & Norvig (2010) AI definitions are cetagorized into four groups. These are: thinking like a human, acting like a human, thinking rationally, and acting rationally. The goal of AI include thinking and acting like a human (Kühl et al, 2020). In other words, AI aims not only to understand intelligence but also to create intelligent beings (Russel & Norvig, 2010). Intelligence is defined as the ability to acquire and apply knowledge and skills while AI is defined as the science of creating artificial entities that from experiences, process and use natural language and develop knowlendge (Balas et al.,2020). AI is seen as the effort to endow computers with human-like characteristics such as perception, association, planning, and reasoning (Boden, 2018).

Human beings have been conducting research on human intelligence and cognitive processes for many years. To this day, human intelligence has not been fully deciphered (Deary et al., 2010). For this reason, the definition of "machines imitating complex human skills" has been seen as a strict definition of AI.

* PhD Student., Ankara University, Faculty of Educational Sciences, Ankara-Türkiye, kubrayilmaz.edu@gmail.com, ORCID ID: 0000-0003-1945-0960
** Prof. Dr., Ankara University, Faculty of Educational Sciences, Ankara-Türkiye, zlfkrdnz@yahoo.com, ORCID ID: 0000-0003-0920-538X

To cite this article:
Yılmaz, K., & Deniz, K. Z. (2024). Natural language processing and machine learning applications for assessment and evaluation in education: opportunities and new approaches. *Journal of Measurement and Evaluation in Education and Psychology, 15*(4), 421-445. https://doi.org/10.21031/epod.1551568

Received: 17.09.2024
Accepted: 28.12.2024

In order to make this definition, it is necessary to identify human-specific skills very well before imitation (Sheikh et al.,2023). In the widest sense, AI is defined as systems designed by humans that interpret data collected from the digital and physical worlds to perform complex tasks given by humans. These systems are established to achieve the best performance in reaching a predetermined goal based on the parameters set by the given data (European Commission, 2018).

This study aims to draw a general framework artificial intelligence (AI), natural language processing (NLP) and machine learning (ML) technologies, to examine the studies on the use of these technologies in education, and to concretise them in the minds of the reader with sample applications that may have potential for the use of these technologies in measurement and evaluation. In line with these aims, it aims to contribute to the field. In this study, topics such as automatic item generation, text visualisation, sentiment analysis, sentence similarity and providing feedback to students are discussed and explained with relevant examples. Firstly, the concept of artificial intelligence will be discussed. It is important to examine the developments in this field from past to present in a historical context.

The Historical Development of AI

When examining the historical development of AI, it will be seen that its foundations were laid in the 1950s. Alan Turing, in his article published in Mind journal, posed the question, "Can machines think?" (Turing, 1950). The most serious response to the question "Can machines think?" was given by McCarthy, Minsky, Rochester, and Shannon (1955) at the Dartmouth Conference held in New Hampshire. This conference is considered a written framework for the concept of "AI." The conference addressed topics that shed light on modern AI, such as automatic computers, programming the use of language, and neural networks.

The years 1950-1970 mark the period when the first research in AI was conducted. During these years, the first AI program, “Logic Theorist,” was developed by Newell and Simon (1956). By 1966, Weizenbaum (1966) developed a program called “Eliza,” which is considered a fundamental work in the field of NLP. In 1969, the first mobile robot capable of perceiving its environment, named “Shakey,” was developed by the Stanford Research Institute (SRI) (Nilsson, 1984). 1970-1980 In this period, which was dubbed as the “AI Winter”, governments reduced funding for AI with the impact of the AI report presented by Lighthill (1973) and developments were interrupted. Reasons such as excessive expectations, technological limitations and lack of data led to the AI winter (Hendler, 2008).

The 1980s-2000s were the period when interest in AI revived and new methods started to be developed. In the 1980s, expert systems had a profound impact on the field of AI, and it was suggested that expert systems could produce solutions to real-world problems. These developments revitalized public interest in AI and raised expectations again (Buchanan & Shortliffe, 1984). After 1990, the fields of ML and data mining gained importance with the increasing amount of data. By the 2010s, deep learning models that can discover complex structures in large data sets and perform operations such as image, video and audio processing have made progress in the field with large data sets and powerful processors (LeCun, Bengio, & Hinton, 2015). Today, AI applications have been used in many areas of life with models such as Gemini developed by Google, GPT-3, GPT-4, GPT 4o developed by Open AI.

At this point, it is important to define the basics of ML and NLP and information about their usage areas in order to ground the subject and to have information about the developments.

Machine Learning (ML)

The term ML was coined by Arthur Samuel in 1959 (Burkov, 2019). ML is based on three concepts: data, model and learning (Deisenroth, Faisal, & Ong, 2020). These systems require large data sets (Kış, 2019). After the data is transferred to the computer environment, models are created in a way that the computer can understand. These models are trained with training data and the accuracy of the trained model is tested with test data. For example, ML algorithms are used in learning fraud detection for credit card transactions and in developing accident prevention systems for cars (Shalev-Shwartz & Ben-David,

2014). It is also used in areas such as face recognition and identification. In ML, decision-making processes are automated after learning based on pre-given samples (Yang & Halim, 2022).

Shalev-Shwartz and Ben-David (2014) explain how the machine learns with an example from natural life. For example, when mice encounter a food whose appearance is different from the previous ones, they eat a small amount, and if the food produces a negative effect, the food is associated with disease. Afterwards, the mice do not eat it when a similar food is encountered. A learning mechanism is at work here. Similarly, when the user marks a mail that falls into the mailbox as 'spam', it informs the AI which mail is 'important' and which mail is 'junk'. When a new e-mail arrives, the machine learns whether to put it in the important folder or the junk folder.

In order for machines to be systems that can think like humans, supervised and unsupervised ML models, regression, classification and clustering are used (Shalev-Shwartz & Ben-David, 2014). ML algorithms consist of four categories developed for different purposes. These are: Supervised Algorithms, Unsupervised Algorithms, Semi-supervised Algorithms, Reinforcement Algorithms (Burkov, 2019).

Supervised ML algorithms are algorithms that require some supervision from the developer. The goal of these algorithms is to predict the target variable using a function defined over a set of independent variables (Burkov, 2019; Mahesh, 2018). Linear regression, logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN) and naive bayes algorithms are examples of supervised ML algorithms (Mahesh, 2018).

Unsupervised ML algorithms are used when the information used to train is not classified or labeled. While there is a goal in supervised learning, there is no goal in unsupervised learning and an inference is reached (Mahesh, 2018). Processes such as clustering, dimensionality reduction, outlier detection are examples of unsupervised ML. Because in these processes, the model works by making inferences from the natural structure of the data and discovering relationships between data instead of pre-labeled data (Burkov, 2019). Semi-supervised ML algorithms use both labeled and unlabeled data for training. They usually use small amounts of labeled data and large amounts of unlabeled data (Chapelle et al., 2006; Burkov, 2019).

Reinforcement ML uses a technique called exploration; the machine interacts with its environment by generating actions, observes the results, and then takes these results into account when performing the next action. The process continues in this way until the algorithm evolves and chooses the right strategy (Mahesh, 2018). Reinforcement learning is used to solve problems with long-term goals where decision-making stages are sequential, such as game playing, resource management, robotics and logistics (Burkov, 2019).

Artificial neural networks, which are inspired by biological neural networks, work similar to the human body's processes of transmitting stimuli to the brain and responding. There are many hidden layers between the input and output layers. In this way, the strength of the network connections determines the output to be transmitted to the next layer by processing the data coming from the input during the learning process of the network, and thus the model becomes capable of making accurate predictions (Yang & Halim, 2022). ML allows us to make various inferences from data by training models with large data sets. NLP is used for processing text data and obtaining meaningful inferences.

Natural Language Processing (NLP)

Language has an important place in human history. People communicate with each other through languages. People's studies on natural languages shed light on today's research on NLP (Oflazer, 2016: Oflazer & Saraçlar, 2018). Developments in information technologies have encouraged people to study languages (Adalı, 2012).

Computers need to use NLP processes in order to understand and communicate with human languages (Şeker, 2015). NLP is a broad set of technologies used to analyze texts semantically and syntactically and to extract meaning (Wijeratne et al., 2009). NLP is a computerized approach to analyzing text. There is no single agreed definition of NLP (Liddy, 2001).

NLP studies and text mining studies are often used together. Text mining studies consist of studies that accept text as a data source. For example, it is used in studies such as classification of texts, extraction of topics from texts, classification, sentiment analysis, text summarization, entity relationship modeling (Şeker, 2015). The branch of science called NLP studies the processing of languages with the help of computers (Adalı, 2012).

One of the most basic concepts of NLP is the concept of corpus. Corpus can be defined as texts written in a language. Corpuses are often used to study language features, train large language models and improve NLP algorithms (Jurafsky & Martin, 2024)**.** Language models are trained with textual data and can learn the structure of the language and language rules with the help of these corpora. Thus, they can make more accurate predictions (Dong, 2023). Turkish has a very rich corpus structure (Sak et al., 2011).

Turkish is in the Turkic languages group of the Altaic language family. Other languages in this language family are Mongolian, Tungusic, Korean and Japanese. For agglutinative languages such as Turkish, there are various difficulties in NLP (Oflazer & Saraçlar, 2018). However, nowadays, various operations can be performed in the field of Turkish NLP with the help of Turkish NLP tools. There are tools developed for Turkish NLP and libraries such as Zemberek-NLP, Turkish Stemmer, TrTokenizer, Mukayese (Merdun et al., 2024; Usta, 2024).

Topics such as NLP and ML are also attracting attention in the field of education and studies are being carried out to integrate them into education (Gierl et al., 2008; Gierl & Lai, 2016, Uysal, 2019; Göloğlu-Demir & Yılmaz, 2018; Mulianingsih et al., 2020; Coelho et al. 2023: Sytnyk & Podlinyayeva, 2024). Other topics to be addressed within the scope of this study are text mining, topic modelling, sentence similarity, student feedback, sentiment analysis, and automatic question generation. See Appendices for all sample applications and code examples.

Text Mining and Topic Modeling

The most common definition of the term data mining is discovering patterns for data (Leskovec et al., 2014). One branch of data mining is text mining. Text mining is concerned with how to determine what the subject of a document is about. Text mining is a method for categorizing texts that contain many topics such as news articles and blog posts into groups according to their topics (Silge & Robinson, 2017). Topic modeling is a method used in NLP applications such as sentiment analysis, document classification, speech recognition, automatic translation. In terms of text mining, topic modeling is based on the bag-of-words assumption (Alghamdi & Alfalqi, 2015).

Topic modeling can provide methods to automatically organize, understand, search and summarize large text data without manual work (Blei et al., 2003). Topic modeling is used to discover patterns of word usage in a document (Alghamdi & Alfalqi, 2015).

Topic modeling is a generative bag-of-words model that learns topics and topic words from frequency measures in texts (Mazidi, 2018). It is a system based on ML and NLP. The LDA model developed by Grun and Hornik (2011) as an R package is used as a topic modeling approach. With the LDA model, subtopics in texts are modeled and thus texts are grouped according to their content similarities. For example, news texts in seven different categories in Turkish (world, economy, culture and arts, health, politics, sports and technology) can be classified under the category they belong to (Yıldırım & Yıldız, 2018). The main point here is that ML learns which category the topics in the text belong to and assigns the texts to the class they belong to.

Text mining is seen as a method that can be used in many areas such as monitoring and evaluating student performance, providing feedback and support to students, and discovering the points where students have difficulty (Ferreira et al., 2020). Word clouds allow some of the findings obtained from text mining to be presented in visualised form. For example; with the help of a word cloud, it is possible to visually see which words are used more by students who answer a question correctly and which words are used more by students who answer incorrectly. In this study, word cloud is introduced with a sample application. See Appendices for application and code examples.

Sentence Similarity

Sentence similarity is one of the most widely used subfields of NLP. It is used to measure the similarity between sentences for tasks such as question answering, information retrieval, summarisation and plagiarism detection (Farouk, 2019). In the sentence similarity approach, words are represented as vectors and the similarities between these vectors are calculated mathematically (Mikolov et al., 2013). Sentence similarity can be interpreted as semantic inferences at the sentence level (Guu et al., 2018). For example, when a customer asks a bank's chatbot a question about a loan, the chatbot matches it with the questions stored in its memory and determines the appropriate response to the customer.

In the sentence similarity approach, the degree of similarity between a reference answer and other answers can be measured. An example of how this method can be used in education is the evaluation of open-ended exam answers by comparing student answers to a reference answer. The closer the similarity score is to 1, the closer the student's answer is to the reference answer and therefore considered correct. For example, Chamidah et al. (2021) the study introduced an essay evaluation system that utilises the similarity between student answers and reference answers using short-answer questions from Indonesia. The method extended the reference answers with synonyms and compared them using Cosine similarity, Jaccard similarity and Dice similarity measures. The questions are categorised into four topics: politics, lifestyle, sport and technology.

In a study developed to evaluate short answers in education, students' answers were compared with reference answers using a semantic similarity measure. The results showed a correlation of 0.70 between manual evaluation and system evaluation. The developed system provides fast and consistent scoring of short answers and significantly supports the reliability of manual scoring (Lubis et al., 2021).

The closer the similarity score is to 1, the closer the student's answer is to the reference answer and therefore considered correct (Wang & Dong, 2020; Chamidah et al., 2021). This method is also seen as a potential tool to detect similarities between student answers and prevent plagiarism. Moreover, this method is seen as a functional tool for measuring language skills and written expression abilities. Providing feedback on students' learning is another important aspect.

Feedback to Students

Feedback is necessary for identifying deficiencies that need to be addressed in education, providing various improvements and developing curricula in this context. The historical development of feedback can be traced back to Thorndike's “Law of Effect” (Lipnevich & Panadero, 2021). Giving feedback to students is important in education. However, giving feedback to each student individually is challenging in terms of time and effort (Cavalcanti et al., 2019).

Effective feedback has some characteristics. Feedback requires that what is expected from the student as a result of the evaluation of the task performed by the student is conveyed to the student in an understandable way. Feedback should be in a way that contributes to the student's learning processes and plays a constructive role. At the same time, it should cover all aspects of the task by focusing on missing gains (Kayalı et al., 2019).

One study is working on a system that gives instant feedback to students. The study is conducted as a pilot study on 800 students studying at a university in India. In this study, it is aimed to give almost real-time feedback to students' writings through the system (Lewkow et al., 2016). In another feedback study, both automatic scoring and giving feedback were studied. The system provides feedback that will allow the student to make the necessary corrections before submitting their work (Woods et al., 2017).

When the number of students is high, it is very difficult to give feedback to each student individually. By using NLP techniques and clustering analysis, common feedback can be given to students who give similar answers. Thus, communication with students will be maintained and time can be saved while doing so. Instead of giving feedback to the students individually, the answers of the students who give similar answers are processed with NLP, converted into mathematical values and clustered according to similarity measures. The same feedback can be given to the student for the answers in the same cluster.

In addition to giving feedback to the student, various improvements can be made in education with the feedback received from the students about the course. With this feedback, students' emotional states can be determined and steps can be taken to improve education accordingly (Kasumba & Neumann, 2024).

A review of the literature reveals various studies in which automated feedback is used in education, with different techniques being developed for this purpose. For example, Lu and Cutumisu (2021) conducted a two-stage study using deep learning approaches and NLP techniques to provide automated written feedback and assessment in education. In their study, three deep learning models (CNN, CNN + LSTM, and CNN + Bi-LSTM) were tested for automated assessment, with the LSTM model achieving the highest accuracy. This model demonstrated an average performance of 0.73 on the Quadratic Weighted Kappa (QWK) metric, which measures alignment with human evaluation.

For the automated feedback stage, the Constrained Generation by Metropolis-Hastings Sampling (CGMH) method was employed to generate contextually appropriate feedback sentences. These sentences were automatically structured based on errors found in students' writing.
In this study, the students were divided into 3 clusters according to the similarities of their answers and the feedbacks written by the researchers in accordance with the clusters were assigned to the cluster they belonged to by the developed system. With this sample application, a basic level of concretisation for automatic feedback in the reader's mind was aimed to be made. See Appendix for the sample application.

Sentiment Analysis

Sentiment analysis is an application of NLP that examines whether individuals' opinions on a topic are positive, negative or neutral. Since the manual construction and validation of a sentiment lexicon is labor- and time-consuming, many studies have explored automated ways of identifying sentiment-related features in text. According to Dong (2023), two different approaches are used for sentiment analysis. One is a rule-based approach and the other is a ML-based approach. In the ML-based approach, emotions are identified with large amounts of text data to train the model. In the rule-based approach, negative words, emotional words, language features are identified within the framework of predetermined rules and emotions in the text are evaluated as 'positive, negative, neutral'. In addition to ML methods, there are methods developed by researchers for sentiment analysis. Hutto and Gilbert (2014) compared ML and VADER (Valence Aware Dictionary and Sentiment Reasoner) methods on 4000 tweets and found that the VADER method performed better. (Sukmana & Rusydiana, 2023).

In the study conducted by Bostancı and Albayrak (2021), sentiment analysis method was used to extract appropriate advertising content for students during the university preference period. The study was conducted on the comments of 82 twitter and 65 facebook users. Student emotions were classified as optimistic, pessimistic, humorous, productive and extraverted. Accordingly, university advertisement posters were designed for the determined emotional states. In their study, Göloğlu-Demir and Yılmaz (2018) calculated the TF-IDF ratios of the 10 most common words containing 10 positive and 10 negative emotions among 36081 words written by 40 participants for 4 days. As a result of the study, it was concluded that the majority of the students had positive feelings about the project.

In recent years, ML and big language models have also been used in studies (Kasumba & Neumann, 2024; Peña-Torres, 2024). Sentiment analysis can be used in the field of education to obtain students' opinions about a course, subject or teaching method and to make various improvements in educational practices (Peña-Torres, 2024). In recent years, sentiment analysis studies in education have become widespread (Sukmana & Rusydiana, 2023; Lin, 2023; Kasumba & Neumann, 2024; Peña-Torres, 2024). Another subject that is becoming widespread in education is text and question generation (Shin, 2021; Kasumba & Neumann, 2024).

Automated Question Generation

Automatic question generation is divided into template-based and non-template-based approaches (Gierl & Lai, 2016). An example of template-based approaches is IGOR, an automatic item generation tool that allows users to generate a variety of test items and can be applied in areas such as mathematics (Gierl et al., 2008). Template-based approaches use auxiliary information such as text, figures and

graphs to generate test items with logical and appropriate values (Singley & Bennett, 2002). Non-template-based approaches produce new texts using NLP techniques. These technologies produce more rational results and add new dimensions to automatic question generation as training data increases. Various tools are being developed to support text generation, one of which is Texar. Texar is a toolkit that can be used in text generation in the field of NLP (Hu et al., 2018). By using neural networks and NLP techniques for automatic question generation, semantic features of texts can be identified and test items can be generated (Shin, 2021).

Today, generative AI tools can be used for this purpose. GPT-3 achieves high success in text generation by processing large amounts of data with 175 billion parameters (Brown et al., 2020). Language models have developed rapidly in recent years. For example, the GPT (Generative Pre-trained Transformer) model introduced by OpenAI is a comprehensive model. GPT-3 adds a new dimension to all the mentioned text generation approaches. For this model, the entire Wikipedia was used as training material. In addition, text data equivalent to 32 times of Wikipedia was also included in the training data of the model. This model was created with 175 billion parameters (Brown et al., 2020).

GPT-3, with its system structure consisting of 175 billion parameters and thousands of introduced texts, can write an article on any topic in seconds, continue any text (in an unprecedented way), and generate both multiple choice and open-ended questions about the text. Machine learning and NLP systems are described as data hungry. The more data they can be trained on, the more rational results they can produce. With the API support provided by OpenAI, researchers can generate various texts in their fields and obtain text-related questions. GPT models can generate previously unseen texts by utilising training data. The most recently optimised GPT model, GPT-4o, is the result of the development of more advanced algorithms by optimising large language models (LLaMEA - Large Language Model Evolutionary Algorithm) (OpenAI, 2023). Another large language model developed by Google is Gemini.

In a study by Zeinalipour et al. (2024), large language models such as GPT-4-Turbo, GPT-3.5-Turbo and Llama were used for automatic question generation in Turkish. The dataset of the study includes various disciplines such as chemistry, biology, geography, philosophy, Turkish language and literature, and history. According to the findings of the study, big language models can be used effectively in the process of creating educational content. In order to provide a broader framework for the use of these technologies in education, the relevant literature has also been examined in education.

The Use of AI Approaches in Education

The intelligence of a learner is often equated with the ability to recall learnt information (Nafea, 2016). This system ignores individual differences, readiness and varying learning speeds. AI applications in education contribute to students' contextual learning and can provide individualised learning experiences for each student (Chaudhry & Kazim, 2021). The application of AI in education has been the focus of academic research for over 30 years (Hamal et al., 2022).

AI technologies can be used to improve learning experiences, improve educational outcomes, develop individualised learning systems to enable students to learn at their own pace, support teachers in material development, etc., and provide instant support to students through gamification, interactive simulations, virtual assistants (Kotlyarova, 2022). In addition, AI applications can be used in many fields such as automating assessment stages, providing instant feedback to students, providing access to space-independent classrooms, medicine, marketing, engineering education, etc. (Sadiku, Ashaolu, Ajayi-Majebi, & Musa, 2021). Tools such as Intelligent Tutoring Systems (ITS) and contextual learning environments (iTalk2Learn and AIDA) are also among the opportunities offered by AI in education (Chaudhry & Kazim, 2021). It is predicted that AI can reduce inequalities among students by providing personalised learning opportunities to individuals (Nkechi et al., 2024). For example, Holstein, McLaren, and Aleven (2018) concluded in their study with 8 teachers and 286 secondary school students in 18 classrooms that it can reduce the gap in learning outcomes between students with different readiness.

It can also be used in areas such as automatic assessment and teacher observation of student progress (Nafea, 2016). Distance education is another area where AI can be used (Coelho et al.,2023). In this context, technologies such as AI, machine learning have a great potential to provide individualised learning experiences in education and increase the efficiency of learning processes and bring new approaches to educators.

For example, these technologies can be utilised in language learning. Perveen (2021) conducted a study on a group of students attending two different courses on English language learning. Word clouds were used in the study. The findings of the study showed that word clouds are a suitable tool for task-based assessment, especially in pre-reading and pre-writing activities.

Another noteworthy area of study is the use of the counterfactual approach to improve the achievement of at-risk students. Cavus and Kuzilek (2024a; 2024b) conducted two studies on counterfactualism. In the first of these studies, counterfactual methods were utilised to increase student achievement in education. These methods are used to provide more accurate counterfactual explanations (what-if scenarios) to students at risk of failure in education. In the study, it is possible to provide meaningful and effective explanations about which factors need to change to increase student achievement. The NICE method was found to be more effective than other methods (Cavus & Kuzilek, 2024a). The second study by Cavus and Kuzilek (2024b) emphasises the importance of the actionability of counterfactual explanations to provide accurate guidance to students at risk of failure.

The aim of the study is to identify specific characteristics that categorise students as at-risk and to show how adjusting for these characteristics can improve their achievement. For this purpose, model-agnostic explanation methods, in particular LIME and SHAP, were used. The results showed that SHAP is more stable and reliable.

The use of counterfactual explanations provides individualised support for students at risk of failing a course (Smith et al., 2022). In a study aiming to understand the impact of counterfactual explanations on student performance, using data from 134 successful and 148 unsuccessful students, it was concluded that the developed system provides individualised recommendations for each student (Tsiakmaki & Ragos, 2021). It is seen that artificial intelligence technologies can be used in subjects such as individual learning, automatic assessment, instant feedback, and counterfactual explanations can be useful to increase student achievement. Young (2024) mentioned the advantages and disadvantages of integrating AI into education. Accordingly, it is stated that AI provides great improvements in education such as personal learning, assessment, creating relevant educational content and virtual reality. All these study examples show that the use of AI in education is effective and beneficial. There is still a lack of research on the inclusion of AI in educational applications. In this study, topics such as automatic item generation, text visualisation, sentiment analysis, sentence similarity and providing feedback to students are discussed and explained with relevant examples. See Appendices for all sample applications and code examples.

Methods

Research Design

This study is designed to draw a general framework for technologies such as artificial intelligence, machine learning and NLP and to examine the use of these technologies in education. The research also aims to concretise the use of these technologies in the minds of the readers by supporting them with sample applications.

This study examines the potential of artificial intelligence, machine learning and NLP technologies to contribute to measurement and evaluation processes in education. Figure 1 shows the flow chart for the literature review and Figure 2 shows the flow chart for the application. Various searches were made in national and international databases (such as scopus, google scholar, national thesis center) to reach the

articles to be analyzed. A total of 90 sources were examined and detailed explanations were provided for the selected examples. The distribution of the sources is as follows:

Table 1.
Distribution of Sources by Category

Main Category	Count
AI	20
AI in Education	16
NLP	12
Sentiment Analysis	7
Question Generation	7
ML	6
Text Mining	6
Feedback	4
Automated Written Scoring	3
Turkish NLP	3
Sentence Similarity	2
GPT-Gemini Models	2
NLP Library	2
Total	90

Search Criteria

The keywords used in the research are: ‘artificial intelligence’, ‘machine learning’, ‘natural language processing’, ‘artificial intelligence in education’. Databases such as ‘Scopus, Google Scholar, National Thesis Centre’ were used in the study. Research tools such as ‘Connected Papers and Typeset’ were also used in the literature review.

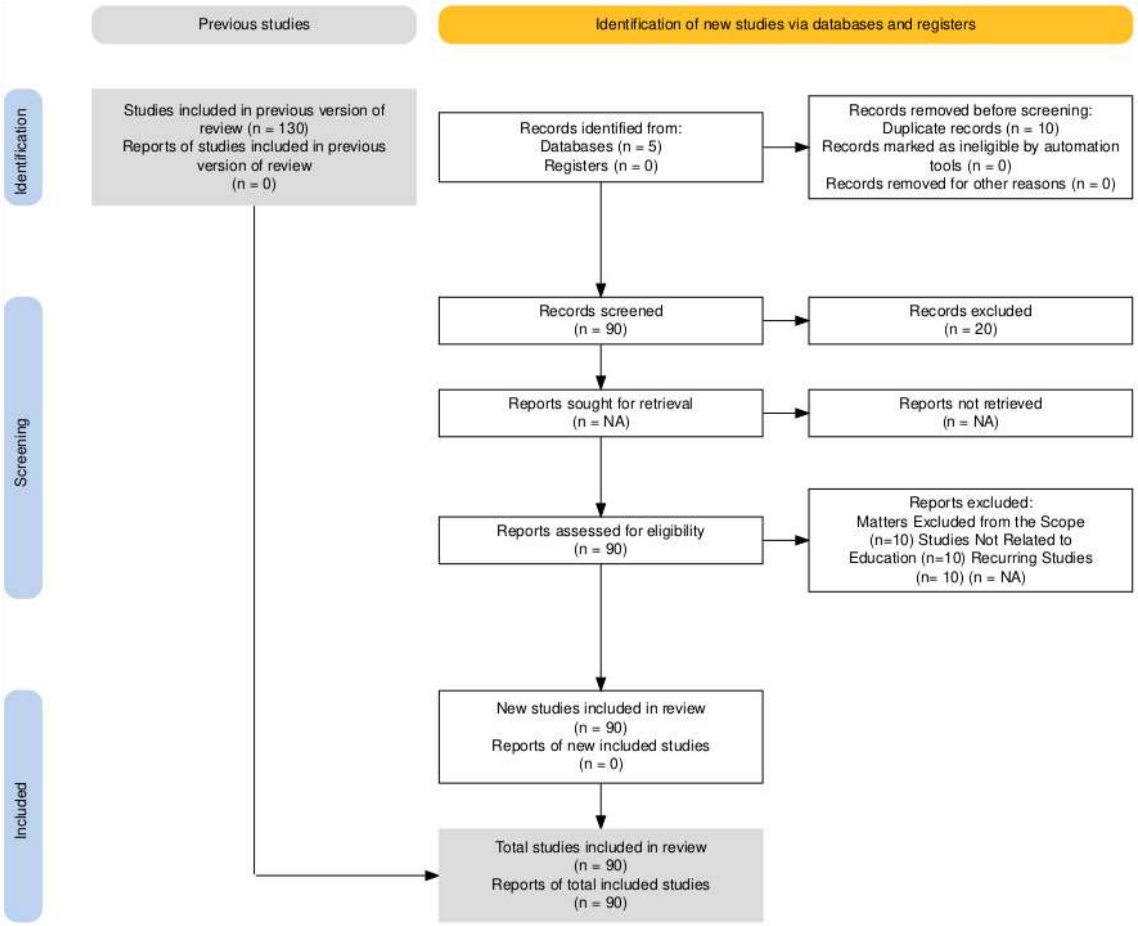
Search Process

90 studies related to the keywords were included in the review, and non-related studies were eliminated. The details of this process are given in the prisma diagram in the following section.

Selection Criteria

In the articles selected for the review, attention was paid to the potential use of artificial intelligence, machine learning and NLP technologies in the field of measurement and evaluation in education and to include application examples. The flow chart of the application steps is as follows:

Figure 1.
Research Process Prisma Flow Diagram (Haddaway, Page, Pritchard and McGuinness, 2022).



Data Analysis

In addition to the literature review, sample applications were also included in the study. The methods and tools used in the study are as follows.

Tools and Methods

Sentence similarity is related to the identification and comparison of text features (Mohler & Mihalcea, 2009). Sentence Transformers model was used for sentence similarity. Turkish Zeyrek (Zeyrek, 2020) and Pandas (McKinney, 2010) libraries were used for sentiment analysis. In addition, Seaborn (Waskom, 2020) and Matplotlib (Hunter, 2007) libraries were used for data visualisation.

In this step, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorisation method was used. In the text vectorisation stage, TF expresses the frequency of occurrence of the word in the document, while IDF measures the ability of the word to distinguish between categories. In this study, texts were converted into numerical values by TF-IDF vectorisation (Shin, 2021; Chen el al., 2016). After this process, student responses were clustered. KMeans algorithm in Python 3.6.11 was used for clustering analysis.

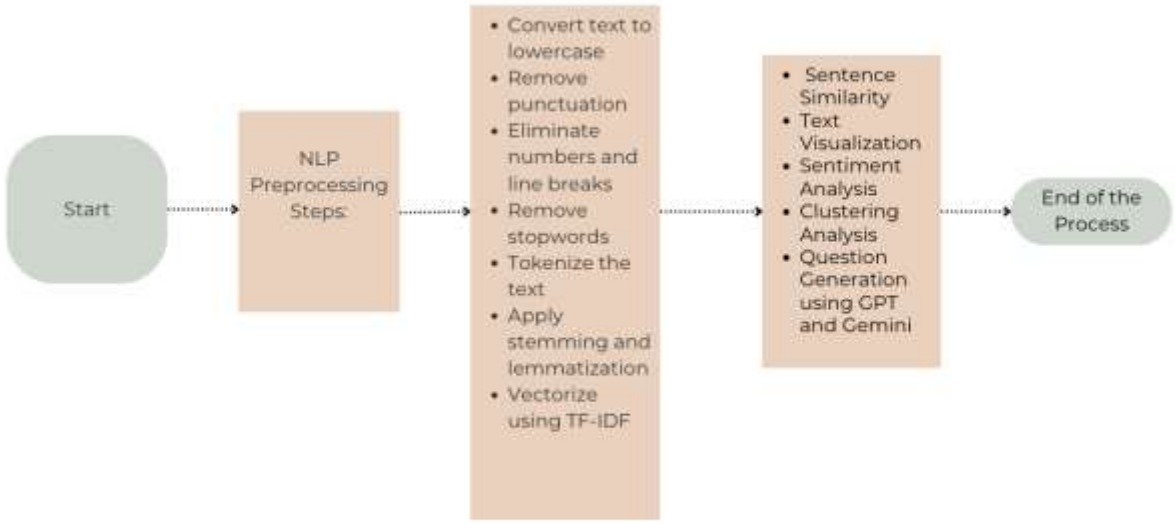
In this step, NLP data preprocessing steps were performed by the researchers. Feedbacks were written by the researchers in accordance with the student responses divided into 3 clusters. For feedback, Scikit-Learn (Pedregosa et al., 2011) and Pandas libraries were used to analyse student responses and assign appropriate feedback to these responses.

Using ChatGPT and GPT-4

In the clustering analysis part, coding assistance was obtained from ChatGPT. In the sentiment analysis stage, student responses were generated with ChatGTP and sentences were grouped according to the appropriate emotion (positive, negative, neutral). Language models such as GPT-4 and Gemini were actively used for question generation.

In this study, data preprocessing steps, the first step of NLP, were performed. Sentence similarity and word cloud operations were coded by the researchers. NLP data preprocessing steps are shown in Figure 2. Sample applications were implemented using Python programming language and these applications are presented in the Appendix with code samples. The flowchart of the application steps is as follows:

Figure 2.
Application Flowchart



Conclusions

Text Visualisation Examples

The use of word clouds helps students to grasp the main theme of the text by highlighting words that occur frequently in sentences. Important findings in the text visualisation literature were presented by Perveen (2021). In this study, an example of a word cloud visualised from the 'Anthem of Independence' was given. A word cloud was created with the most frequently mentioned words in the National Anthem. In addition, word cloud examples containing the correct and incorrect answers given by the students to the question 'What is the meaning and importance of the National Anthem?' were created using GPT. Especially in the 'İstiklal Marşı' example, it is foreseen that it can provide an opportunity to evaluate students' understanding of the main themes of the anthem such as patriotism, freedom and unity. Visual objects can be useful in terms of memorisation. The words that stand out in students' correct and incorrect answers can also be evaluated.

Sentence Similarity Examples

Two examples implemented in Python programming language are given in the Appendix. In the first example, the sentence similarity score is calculated as 0.77. In the other example, the similarity score

between two sentences is calculated as 0.91. A similarity score of 0.77 indicates a moderate level of similarity between the two sentences, while a higher score of 0.91 indicates that the sentences share almost the same meaning.

The results obtained in this study (0.77; 0.91) are higher than the results obtained by Lubis et al., (2021) (0.70) for both sentence pairs. When used in automated assessment systems, these scores reflect how close the student's answer is to the model answer, allowing an objective assessment to be made. This method can be used to help graders in the evaluation of open-ended questions and to speed up the process.

Student Feedback Example

In the example below, the text and responses were generated using GPT-3.5. In the analysis of the text, Turkish NLP techniques and clustering analysis, an unsupervised ML model, were used in Python programming language. Student responses were categorised into three clusters according to their similarities. Predetermined feedbacks for each cluster were automatically added next to the responses in that cluster. In the literature, Lu and Cutumisu (2021) conducted more extensive studies on automatic written feedback in education. In this study, a basic example for automatic feedback is presented and an example of the use of both ML and NLP is created.

Sentiment Analysis Example

In this study, a sample application was made at a basic level. Student views on a mathematics lesson created using GPT-4o, a code sample for sentiment analysis and a pie chart of student views were created. Accordingly, 4 out of 10 students expressed positive, 4 negative and 2 undecided opinions about the mathematics lesson. Research on sentiment analysis emphasises the potential of these technologies in education (Kort et al., 2001; Peña-Torres, 2024). It is seen that the use of sentiment analysis in education will be beneficial in studies conducted on real students with larger data sets.

Item Generation Example

In studies on question generation (Gierl et al., 2008; Gierl & Lai, 2016; Shin, 2021), it is seen that models developed by researchers are used. However, studies on generating questions with GPT have also increased in recent years (Smith et al., 2024; Berger et al., 2024). In this study, a text was created with GPT4 in the question generation phase and open-ended, multiple-choice questions were generated based on this text. Similarly, both open-ended and multiple-choice questions were generated with Gemini. In future studies, questions created with both tools can be applied to real students and the results can be evaluated. See Appendices for all sample applications and code examples.

Discussion and Suggestions

This study addresses the historical development of AI, ML, NLP techniques, and their implications for education, including both a literature review and practical applications. The results indicate that approaches like ML and NLP are applicable in educational settings (Zeinalipour et al., 2024; Smith et al., 2024; Berger et al., 2024). Student feedback plays a critical role in improving the learning process in education; Lu and Cutumisu (2021) and Kasumba and Neumann (2024) provide significant findings on this topic. Through counterfactual validity, factors that contribute to improving student success can be focused on (Cavus & Kuzilek, 2024a). In addition to traditional teaching methods, new approaches such as ML and NLP can contribute to students' learning processes by supporting each other (Nafea, 2016).

If we consider the points that all these applications will contribute to the field of measurement and evaluation in education; personalised learning environments will provide individuals with the opportunity to evaluate themselves and progress at their own pace. With tests adapted to the individual, the individual will be evaluated with questions appropriate to his/her own level, which will allow each student to catch his/her own success scale without ignoring individual differences. This has the potential

to contribute to the constructivist approach. These tools will also save time for question writers and practitioners and can be used as an auxiliary tool in the process of preparing questions suitable for each level. Of course, the questions developed with these tools need to be tested on appropriate samples and used in a controlled manner by humans. Thanks to automatic scoring and feedback, the use of open-ended items will become widespread, and more information can be obtained in the measurement of high-level cognitive skills by conducting an interactive process with the student. At the same time, it will contribute to the formative assessment approach and prepare the ground for making necessary improvements in education in the light of this feedback from students. In addition to feedback, emotion analysis, which is another issue addressed in this study, will support these improvements by enabling more information to be obtained about the student in accordance with the formative assessment approach.

Text visualisation can be interesting for students. In addition, it is foreseen that interactive environments such as simulation, virtual reality, game-based learning that can attract students' interest will also be included in learning and assessment processes. AI technologies have the potential to bring very useful innovations for individuals with special needs. For example, it is predicted that virtual assistants can be designed to provide reading support to students with dyslexia.

AI technologies can be easily used both in K-12 and higher education when the necessary infrastructures are provided. While simpler, easy-to-apply and easy-to-use tools are integrated into educational processes at K-12 level, it is predicted that more complex structures and advanced technologies can be used at higher education level. Scaling issues can be addressed by using open source materials at both levels and developing modular systems suitable for each level. The integration of these tools into education can be tested with pilot applications for both levels of education and the results can be evaluated.

More detailed information about students can be obtained by conducting various studies (ML, NLP) on student data obtained from learning management systems (LMS). In the light of these data, learning environments suitable for students can be designed. Customised chatbots can be developed where students can ask questions on any subject at any time.

All these innovations bring with them ethical and security issues. Young (2024) addressed these issues as data privacy and security, bias and fairness, job loss and equality in education, accessibility and inclusiveness. At this point, ethical principles and guidelines regarding the use of AI in education should be determined, copy detection software should be developed, data privacy and security should be ensured. Accountability and necessary transparency should be provided on how the systems work. At the same time, educator training should be emphasised at the point of AI and educators should be given competence in these issues.

In conclusion, it is evident that the use of these technologies in assessment and evaluation in education has been increasing and holds potential for enhancing student success and improving the quality of education. Future research should focus on further studies regarding the integration of these technologies into assessment and evaluation in education.

The sample questions and student answers in this study were generated with generative artificial intelligence tools such as GPT and Gemini. In future studies, it is suggested that the questions generated with these tools should be applied on real student sets and necessary psychometric studies should be carried out.

In this study, an example of automatic feedback to students was given. In future studies, feedback from students about the courses can be processed and evaluated what kind of improvements can be made in this direction. Generalisability of these technologies can be ensured with studies conducted on different languages.

Declarations

Gen-AI Use: The authors of this article declare that Gen-AI tools have NOT been used in any capacity for content creation in this work.
Author Contribution: The first author led the study and contributed to conceptualization, methodology, data modeling, analysis, and visualization, interpretation, and writing. All the other authors played critical roles in shaping the study by contributing to concept, methodology, interpretation, or revision.
Conflict of Interest: No potential conflict of interest was reported by the authors.
Consent to Publish: Written consent was sought from each author to publish the manuscript.
Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Adalı, E. (2012). Doğal dil işleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5*(2). https://dergipark.org.tr/tr/pub/tbbmd/issue/22245/238797

Alghamdi, R., & Alfalqi, K. (2015). A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications (IJACSA), 6*(1), 147-153.

Balas, V. E., Kumar, R., & Srivastava, R. (Eds.). (2020). *Recent trends and advances in AI and internet of things*. Springer Nature. https://doi.org/10.1007/978-3-030-32644-9.

Başarır, L. (2022). Modelling AI in architectural education. *Gazi University Journal of Science, 35*(4), 1260-1278. https://doi.org/10.35378/gujs.967981

Berger, M., Kinsley, A., & Chawla, S. (2024). A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. *arXiv preprint arXiv:2401.07098*. https://arxiv.org/abs/2401.07098

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of ML Research, 3*, 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Boden, M. A. (2018). What is AI? In *AI: A Very Short Introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780199602919.003.0001

Bostancı, B., & Albayrak, A. (2021). Duygu Analizi İle Kişiye Özel İçerik Önermek. Veri Bilimi, 4(1), 53-60. https://dergipark.org.tr/tr/pub/veri/issue/59505/777675#article_cite

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. https://arxiv.org/abs/2005.14165

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems*. Addison-Wesley.

Burkov, A. (2019). *The hundred-page machine learning book.* Andriy Burkov.

Cavalcanti, A. P., Ferreira Leite de Mello, R., Rolim, V., André, M., Freitas, F., & Gaševic, D. (2019). An analysis of the use of good feedback practices in online learning courses. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)* (pp. 153-157). Maceio, Brazil. https://doi.org/10.1109/ICALT.2019.00061

Cavus, M., & Kuzilek, J. (2024a). *An effect analysis of the balancing techniques on the counterfactual explanations of student success prediction models*. arXiv preprint arXiv:2408.00676.

Cavus, M., & Kuzilek, J. (2024b). *The Actionable Explanations for Student Success Prediction Models: A Benchmark Study on the Quality of Counterfactual Methods*. arXiv preprint arXiv:2405.14016.

Chamidah, N., Santoni, M. M., Irmanda, H. N., Astriratma, R., Tua, L. M. & Yuniati, T. (2021). Word Expansion using Synonyms in Indonesian Short Essay Auto Scoring. *International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS).* doi:10.1109/ICIMCIS53775.2021.9699374

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-Supervised Learning*. MIT Press. Retrieved from http://www.acad.bg/ebook/ml/MITPress-%20SemiSupervised%20Learning.pdf

Chaudhry, M. A., & Kazim, E. (2021). AI in Education (AIEd): a high-level academic and industry note 2021. *AI and Ethics*, 2(157-165). https://link.springer.com/article/10.1007/s43681-021-00074-z

Chen, J., Chen, C., & Liang, Z. (2016). Optimized TF-IDF algorithm with the adaptive weight of position of word. In *2nd International Conference on AI and Industrial Engineering (AIIE2016) Advances in Intelligent Systems Research* (Vol. 133).

Coelho, A. M. L., da Silva, H. F., da Silva, L. A. C., Andrade, M. E., & Rodrigues, R. G. da S. (2023). Inteligência artificial: Suas vantagens e limites em cursos à distância. *Revista Ilustração*, 4(2), 23-27. https://doi.org/10.46550/ilustracao.v4i2.150

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience, 11*(3), 201-211. https://doi.org/10.1038/nrn2793

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for ML*. Cambridge University Press. https://mml-book.com

Dong, J. (2023). NLP pretraining language model for computer intelligent recognition technology. *ACM Transactions on Asian and Low-Resource Language Information Processing*. Retrieved from : https://dl.acm.org/doi/pdf/10.1145/3605210

European Commission. (2018). *Definition of AI*. High-Level Expert Group on AI (AI HLEG). Retrieved from https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf.

Farouk, M. (2019). Measuring sentences similarity: A survey. *Indian Journal of Science and Technology, 12*(25). https://doi.org/10.17485/ijst/2019/v12i25/143977

Ferreira, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2020). Text Mining in Education. Retrieved from: https://arxiv.org/pdf/2403.00769

Gierl, M. J., & Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice, 35*(4), 6-20. Retrieved from https://doi.org/10.1111/emip.12136.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning and Assessment, 7*(2).

Göloğlu Demir, C., & Yılmaz, H. (2018). Sınıf dışı eğitim faaliyetlerinin öğrencilerin bilim ve teknolojiye yönelik tutumlarına etkisi ve duygu analizi. *İnsan ve Toplum Bilimleri Araştırmaları Dergisi, 7*(5), 101-116. https://doi.org/10.15869/itobiad.483404

Grun, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. Retrieved from https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

Guu, H., Hashimoto, T. B., & Oren, Y. (2018). Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics, 6*, 437-450. https://doi.org/10.1162/tacl_a_00030

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis Campbell Systematic Reviews, 18, e1230. https://doi.org/10.1002/cl2.1230

Hamal, O., El Faddouli, N., Alaoui Harouni, M. H., & Lu, J. (2022). AI in Education. *Sustainability, 14*(2862). https://doi.org/10.3390/su14052862

Hendler, J. (2008). Avoiding another AI winter. *IEEE Intelligent Systems, 23*(2), 2-4.

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. *AI in Education: 19th International Conference*, 154-168. https://doi.org/10.1007/978-3-319-93843-1_12

Hu, Z., Yang, Z., Shi, H., Tan, B., Zhao, T., He, J., Liang, X., Wang, W., Yu, X., Wang, D., Qin, L., Ma, X., Liu, H., Singh, D., Zhu, W., & Xing, E. P. (2018). Texar: A modularized, versatile, and extensible toolbox for text generation. *Proceedings of Workshop for NLP Open Source Software*, 13-22. https://doi.org/10.18653/v1/W18-2503

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In E. Adar & P. Resnick (Eds.), *Proceedings of the eighth international AAAI*

conference on weblogs and social media. Retrieved from https://ojs.aaai.org/index.php/ICWSM/issue/view/274

İlikhan, S., Özer, M., Tanberkan, H., & Bozkurt, V. (2024). How to mitigate the risks of deployment of AI in medicine? *Turkish Journal of Medical Sciences*, 54(3), 483-492. https://doi.org/10.55730/1300-0144.5814

Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing: An introduction to NLP, computational linguistics, and speech recognition*.

Kasumba, R., & Neumman, M. (2024). Practical Sentiment Analysis for Education: The Power of Student Crowdsourcing. *Proceedings of the AAAI Conference on AI*, *38*(21), 23110-23118. https://doi.org/10.1609/aaai.v38i21.30356

Kayalı, B., Balat, Ş., Kurşun, E., & Karaman, S. (2019). Lisansüstü eğitimde etkili ve nitelikli geribildirim. *Journal of Instructional Technologies & Teacher Education, 1*(8), 10-20.

Kış, A. (2019). Eğitimde yapay zeka. In *14. Uluslararası Eğitim Yönetimi Kongresi Tam Metin Bildiri Kitabı*.

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy—Building a learning companion. *Proceedings IEEE International Conference on Advanced Learning Technologies*, 43-46. https://doi.org/10.1109/ICALT.2001.943850

Kotlyarova, I. O. (2022). AI technologies in education. *Bulletin of the South*

Kühl, N., Goutier, M., Hirt, R., & Satzger, G. (2020). Machine learning in artificial intelligence: Towards a common understanding. *arXiv:2004.04686 [cs.LG]*. https://doi.org/10.48550/arXiv.2004.04686

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press. Retrieved from http://infolab.stanford.edu/~ullman/mmds/book0n.pdf

Lewkow, N., Kode, S., Feild, J., Zimmerman, N., Riedesel, M., Essa, A., Boulanger, D., Seanosky, J., Kumar, V., & Kinshuk. (2016). A scalable learning analytics platform for automated writing feedback. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*. https://doi.org/10.1145/2876034.2893380

Liddy, E. D. (2001). NLP. In *Encyclopedia of library and information science* (2nd ed.). Marcel Decker, Inc. Retrieved from https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub

Lin, F. (2023). Sentiment analysis in online education: An analytical approach and application. *Proceedings of the 2023 International Conference on ML and Automation*. Retrieved from https://doi.org/10.54254/2755-2721/33/20230225.

Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education, 6*. https://doi.org/10.3389/feduc.2021.720195

Lighthill, J. (1973). *Artificial intelligence: A general survey*. In *Artificial intelligence: A paper symposium* (pp. 1-77). Science Research Council. https://www.aiai.ed.ac.uk/events/lighthill1973/lighthill.pdf

Lu, C., & Cutumisu, M. (2021). *Integrating deep learning into an automated feedback generation system for automated essay scoring*. Paper presented at the International Conference on Educational Data Mining (EDM). International Educational Data Mining Society. https://files.eric.ed.gov/fulltext/ED615567.pdf

Lubis, F. F., Mutaqin, A. P., Waskita, D., Sulistyaningtyas, T., Arman, A. A., & Rosmansyah, Y. (2021). Automated short-answer grading using semantic similarity based on word embedding. *International Journal of Technology, 12*(3), 571-581.

Mahesh, B. (2018). ML algorithms - A review. *International Journal of Science and Research (IJSR), 9*(1). https://www.ijsr.net/archive/v9i1/ART20203995.pdf

Mazidi, K. (2018). Automatic Question Generation From Passages. In A. Gelbukh (Ed.), *CICLing 2017, LNCS 10762* (pp. 655-665). Springer. https://doi.org/10.1007/978-3-319-77116-8_49

Merdun, G., Okçular, E., Altınok, D., & Akkurt, F. (2024). *Turkish NLP Resources*. GitHub repository. Retrieved July 10, 2024, from https://github.com/agmmnn/turkish-nlp-resources

McCarthy, J. (2007). What is AI? Computer Science Department, Stanford University. Retrieved from http://www-formal.stanford.edu/jmc/whatisai.pdf

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1956). A proposal for the Dartmouth summer research project on AI. Retrieved from http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McKinney, W. (2010). Data analysis in Python. *Proceedings of the 9th Python in Science Conference* (pp. 51-56).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mulianingsih, F., Anwar, K., Shintasiwi, F. A., & Rahma, A. J. (2020). AI dengan Pembentukan Nilai dan Karakter di Bidang Pendidikan. Ijtimaiya: Journal of Social Science Teaching, 4(2), 148-154. Retrieved from http://journal.stainkudus.ac.id/index.php/Ijtimaia

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 567–575. Association for Computational Linguistics.

Nafea, I. T. (2016). ML in educational technology. In ML- Advanced techniques and emerging applications. IntechOpen. http://dx.doi.org/10.5772/intechopen.72906

Newell, A., & Simon, H. A. (1956). The logic theory machine. *IRE Transactions on Information Theory, 2*(3), 61-79.

Nilsson, N. J. (1984). Shakey the robot. SRI International. Retrieved from https://www.sri.com/publication/artificial-intelligence-pubs/shakey-the-robot-pub/

Nkechi, A. A., Ojo, A. O., & Eneh, O. A. (2024). Impact of AI in Achieving Quality Education. IntechOpen. https://doi.org/10.5772/intechopen.1004871

Oflazer, K. (2016). Türkçe ve doğal dil işleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5*(2). https://dergipark.org.tr/tr/pub/tbbmd/issue/22245/238795

Oflazer, K., & Saraçlar, M. (2018). *Turkish NLP*. Springer.

OpenAI. (2023). GPT-4 Turbo and GPT-4. OpenAI. https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Peña-Torres, J. A. (2024). Towards an improved teaching practice using sentiment analysis in student evaluation. *Ingeniería y Competitividad, 26*(2), e-21013759. Retrieved from: https://www.researchgate.net/publication/381598280_Towards_an_improved_of_teaching_practice_using_Sentiment_Analysis_in_Student_Evaluation.

Perveen, A. (2021). Use of word clouds for task-based assessment in asynchronous e-language learning. *MEXTESOL Journal, 45*(2).

Ramachandran, D., & Rana, R. S. (2024). AI for legal system: Jurisprudence in the digital age. *International Journal of Advanced Academic Studies, 6*(5), 03-13. https://doi.org/10.33545/27068919.2024.v6.i5a.1158

Russell, S., & Norvig, P. (2010). *AI: A Modern Approach* (3rd ed.). Pearson.

Sadiku, M. N. O., Ashaolu, T. J., Ajayi-Majebi, A., & Musa, S. M. (2021). AI in education. *International Journal of Scientific Advances, 2*(1), 1-11. https://typeset.io/pdf/artificial-intelligence-in-education-5ggabmq2kf.pdf

Sak, H., Güngör, T. & Saraçlar, M. (2011) Resources for Turkish morphological processing. *Lang Resources & Evaluation* **45**, 249–261. https://doi.org/10.1007/s10579-010-9128-6

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107298019

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology*. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-21448-6

Shin, E. (2021). *Automated item generation by combining the non-template and template-based approaches to generate reading inference test items* (Doctoral dissertation, University of Alberta). Department of Educational Psychology.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Routledge. https://doi.org/10.4324/9781410602145

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media.

Smith, B. I., Chimedza, C., & Bührmann, J. H. (2022). Individualized help for at-risk students using model-agnostic and counterfactual explanations. *Educational and Information Technologies, 27*(2), 1539–1558. https://doi.org/10.1007/s10639-021-10661-6

Smith, J., Li, H., & Patel, R. (2024). Automated generation of multiple-choice cloze questions for assessing English vocabulary using GPT-turbo 3.5. *arXiv preprint arXiv:2403.02078*. https://arxiv.org/abs/2403.02078

Sukmana, R., & Rusydiana, A. S. (2023). Social media sentiment analysis on waqf and education. *Islamic Marketing Review*, 2(2). Retrieved from http://journals.smartinsight.id/index.php/IMR

Sytnyk, L., & Podlinyayeva, O. (2024). AI in education: Main possibilities and challenges. In Proceedings of the 8th International Scientific and Practical Conference "International Scientific Discussion: Problems, Tasks and Prospects" (pp. 569-579). Brighton, United Kingdom. https://doi.org/10.51582/interconf.19-20.05.2024.058

Şeker, S. E. (2015). Metin madenciliği (Text mining). *YBS Ansiklopedi, 2*(3). https://ybsansiklopedi.com/wp-content/uploads/2015/08/MetinMadenciligi30_32.pdf

Tsiakmaki, M., & Ragos, O. (2021). A case study of interpretable counterfactual explanations for the task of predicting student academic performance. *2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, 120-125. https://doi.org/10.1109/CSCC53858.2021.00029

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460. Retrieved From: https://phil415.pbworks.com/f/TuringComputing.pdf

Ural State University. Ser. Education. Educational Sciences, 14(3), 69-82. https://vestnik.susu.ru/ped/article/view/12330

Usta, Y. (2024). *Awesome Turkish NLP*. GitHub repository. Retrieved July 10, 2024, from https://github.com/yusufusta/awesome-turkish-nlp

Uysal, İ. (2019). *Açık uçlu maddelerde otomatik puanlamanın güvenirliği ve test eşitleme hatalarına etkisi* (Doktora tezi, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü). YÖK Ulusal Tez Merkezi. https://tez.yok.gov.tr/UlusalTezMerkezi/tezDetay.jsp?id=55GArTnn6vLwQ3HOxnwo_w&no=gj27xzLBIdoGFgSUJzjT6Q

Wang, J., & Dong, Y. (2020). Measurement of Text Similarity: A Survey. *Information, 11*(9), 421. doi:10.3390/info11090421

Waskom, M. (2020). *Seaborn: Statistical data visualization*. Erişim adresi: https://seaborn.pydata.org

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Wijeratne, Y., Silva, N., & Shanmugarajah, Y. (2009). NLP for government: Problems and potential. Retrieved from https://lirneasia.net/wp-content/uploads/2019/04/Natural_Language_Processing_for_Government__Problems_and_Potential.pdf

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017). Beyond Automated Essay Scoring: Forecasting and Improving Outcomes in Middle and High School Writing. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)* (pp. 2071-2080). ACM. https://doi.org/10.1145/3097983.3098160

Yang, A., & Halim, S. (2022). Natural language generation using ML techniques. *Journal of Student Research, 11*(2). Retrieved from https://typeset.io/papers/natural-language-generation-using-machine-learning-1itplfnn.

Yıldırım, S., & Yıldız, T. (2018). A Comparison of Different Approaches to Document Representation in Turkish Language. *Journal of Natural and Applied Sciences, 22*(2), 569-576. https://doi.org/10.19113/sdufbed.15893

Young, J. (2024). The rise of AI in education. *International Journal of Innovative Research & Development, 13*(2), 74. https://www.internationaljournalcorner.com/index.php/ijird_ojs/article/view/173518/118319

Zeinalipour, K., Keptiğ, Y. G., Maggini, M., & Gori, M. (2024). Automating Turkish educational quiz generation using large language models. https://doi.org/10.48550/arXiv.2406.03397

Zeyrek, A. (2020). *Zeyrek: Morphological analysis for Turkish*. GitHub. https://github.com/ahmetaa/zeyrek

Appendix

Appendix 1.

Examples of Text Visualization

Sample application and code example for text visualisation examples are given below.


```
# Kelime listelerini oluşturma
dogru_kelimeler = set(dogru_cevaplar.split())
yanlis_kelimeler = set(yanlis_cevaplar.split())

# Ortak kelimeleri çıkartma
dogru_kelimeler = dogru_kelimeler - yanlis_kelimeler
yanlis_kelimeler = yanlis_kelimeler - dogru_kelimeler

# Güncellenmiş metinler
dogru_cevaplar = ' '.join(dogru_kelimeler)
yanlis_cevaplar = ' '.join(yanlis_kelimeler)
```

Appendix 2.

Examples of Sentence Similarity

Examples of sentence similarity sample application and code example are given below.

```
cümle1 = "İstanbul 1453 yılında fetholundu."
cümle2 ="Fatih Sultan Mehmet İstanbulu 1453 yılında aldı."
benzerlik oranı:([[0.7761]])
cümle1 = "Atatürk 19 Mayıs 1919'da Samsuna'a ayak baktı."
cümle2 ="19 Mayıs 1919 Atatürk'ün Samsun'a gittiği tarihtir."
benzerlik oranı:([[0.9192]])
```

Appendix 3.

Examples of Sentence Similarity

Student Feedback sample text, application and code example are given below.

Sample Text

Çocukluk arkadaşlarının hafızasında kalan anılar, zamanla solmayan nadir hazinelerdir. Oyunlar, gülüşmeler, hatta küçük kavgalar bile yıllar geçse de unutulmaz. Ahmet ve Mehmet, çocukluklarını aynı mahallede, aynı sokakta geçirmiş iki dosttu. İki arkadaş, her akşam saatlerce sokakta top oynar, macera ararlardı. Bir gün, kocaman bir ağaç gördüler. Gövdesi sağlam, dalları uzanmıştı gökyüzüne. Ahmet, hemen tırmanmaya başladı. Mehmet, cesareti topladı ve arkadaşının ardından ağaca tırmandı. Birlikte tepesine çıktıklarında, etraflarını seyrettiler. Küçük mahalleleri, yeşillikler içindeki parkı, uzaklarda görünen caminin minaresini görebiliyorlardı. O an, ikisi de birbirine gülümsedi. Bu anı, yıllar geçse de unutulmayacak anılardan biri olacaktı.

Question: Metinde Ahmet ve Mehmet'in en unutulmaz anısını sizce ne yaratmış olabilir?

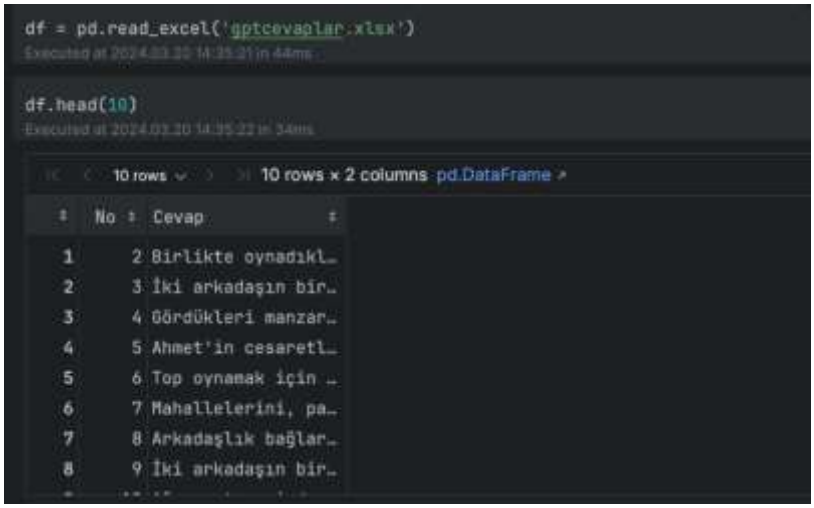
Appendix Table 1.

Example of Student Responses and Feedback

Response	Cluster	Feedback
1. Ağacın tepesindeki sessizliğin ve huzurun tadını çıkarmaları.	2	Tebrikler. Daha fazla okuma çalışması yaparak yorumlama yeteneğini artırabilirsin. Bu noktada okuman için .... kitaplarını önerebilirim.
2. Ahmet'in ağacın tepesinden bulutların şekillerini tahmin etmeye çalışması ve Mehmet'in onunla bakması.	1	Metni tekrar okumanı ve yeni çıkarımlar yapmanı önerebilirim. Konuyu farklı yönleriyle ele almanın yorumlama yeteneğini geliştireceğini umuyorum.
3. İki arkadaşın ağacın tepesindeyken birlikte yıldızları saymaları ve hangi yıldızın hangi burcu temsil ettiğini konuşmaları.	2	Tebrikler. Daha fazla okuma çalışması yaparak yorumlama yeteneğini artırabilirsin. Bu noktada okuman için .... kitaplarını önerebilirim.
4. Ağacın tepesinde otururken Ahmet'in eline düşen yaprağı yakalamaya çalışması ve Mehmet'in ona yardım etmesi.	1	Metni tekrar okumanı ve yeni çıkarımlar yapmanı önerebilirim. Konuyu farklı yönleriyle ele almanın yorumlama yeteneğini geliştireceğini umuyorum.
5. İki arkadaşın ağacın tepesinden çevredeki ağaçlardaki kuşların cinslerini tahmin etmeye çalışmaları.	2	Tebrikler. Daha fazla okuma çalışması yaparak yorumlama yeteneğini artırabilirsin. Bu noktada okuman için .... kitaplarını önerebilirim.
6. Tepeden gördükleri manzarayı birlikte betimleyerek bir hikaye oluşturmaya çalışmaları.	2	Tebrikler. Daha fazla okuma çalışması yaparak yorumlama yeteneğini artırabilirsin. Bu noktada okuman için .... kitaplarını önerebilirim.
7. Ahmet'in ağacın tepesinden aşağıya bir şey atmaya cesaret edememesi ve Mehmet'in onu desteklemesi.	1	Metni tekrar okumanı ve yeni çıkarımlar yapmanı önerebilirim. Konuyu farklı yönleriyle ele almanın yorumlama yeteneğini geliştireceğini umuyorum.
8. İki arkadaşın ağaçta otururken geçmişte yaşadıkları komik anıları hatırlamaları.	3	Keyifli bir yorum getirmişsin. Konuya hakim görünüyorsun. İnsan yaşam boyu öğrencidir. Yeni şeyler öğrenmekten hep keyif alman dileğiyle.

9. İki arkadaşın ağaçta otururken bulutların hareketini izleyerek hayal kurmaları.	3	Keyifli bir yorum getirmişsin. Konuya hakim görünüyorsun. İnsan yaşam boyu öğrencidir. Yeni şeyler öğrenmekten hep keyif alman dileğiyle.

In this study, feedback was generated by clustering responses based on answer similarities.

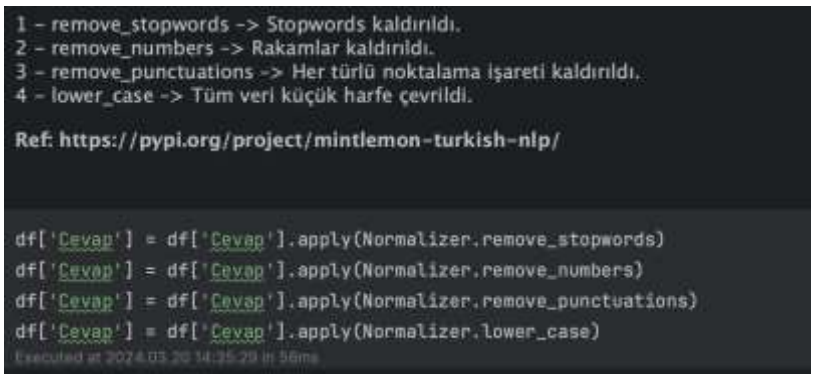


```
df = pd.read_excel('gptcevaplar.xlsx')
```

```
df.head(10)
```

10 rows × 2 columns pd.DataFrame

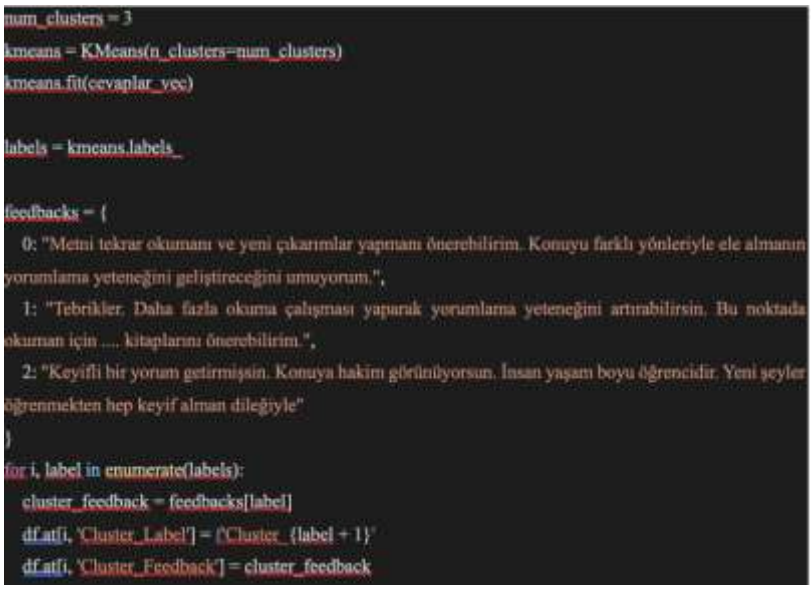
	No	Cevap
1	2	Birlikte oynadıkl…
2	3	İki arkadaşın bir…
3	4	Gördükleri manzar…
4	5	Ahmet'in cesaretl…
5	6	Top oynamak için …
6	7	Mahallelerini, pa…
7	8	Arkadaşlık bağlar…
8	9	İki arkadaşın bir…



```
1 - remove_stopwords -> Stopwords kaldırıldı.
2 - remove_numbers -> Rakamlar kaldırıldı.
3 - remove_punctuations -> Her türlü noktalama işareti kaldırıldı.
4 - lower_case -> Tüm veri küçük harfe çevrildi.

Ref: https://pypi.org/project/mintlemon-turkish-nlp/
```

```
df['Cevap'] = df['Cevap'].apply(Normalizer.remove_stopwords)
df['Cevap'] = df['Cevap'].apply(Normalizer.remove_numbers)
df['Cevap'] = df['Cevap'].apply(Normalizer.remove_punctuations)
df['Cevap'] = df['Cevap'].apply(Normalizer.lower_case)
```



```
num_clusters = 3
kmeans = KMeans(n_clusters=num_clusters)
kmeans.fit(cevaplar_vec)

labels = kmeans.labels_

feedbacks = {
  0: "Metni tekrar okumanı ve yeni çıkarımlar yapmanı önerebilirim. Konuyu farklı yönleriyle ele almanın
yorumlama yeteneğini geliştireceğini umuyorum.",
  1: "Tebrikler. Daha fazla okuma çalışması yaparak yorumlama yeteneğini artırabilirsin. Bu noktada
okuman için .... kitaplarını önerebilirim.",
  2: "Keyifli bir yorum getirmişsin. Konuya hakim görünüyorsun. İnsan yaşam boyu öğrencidir. Yeni şeyler
öğrenmekten hep keyif alman dileğiyle"
}
for i, label in enumerate(labels):
  cluster_feedback = feedbacks[label]
  df.at[i, 'Cluster_Label'] = f'Cluster_{label + 1}'
  df.at[i, 'Cluster_Feedback'] = cluster_feedback
```

Appendix 4.

Example of Sentiment Analysis

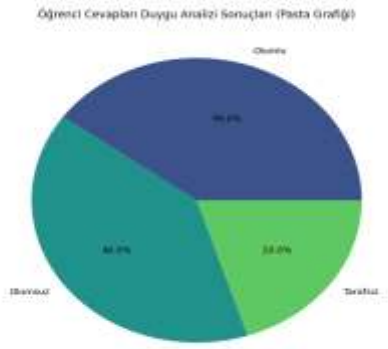
An example of the application and code for emotion analysis is given below.

Appendix Table 2.

Sentiment Analysis Results

Response	Sentiment Analysis Result
Matematik dersini çok seviyorum çünkü problem çözmek bana zevk veriyor.	Olumlu
Matematik dersi benim için zorlayıcı ve sıkıcı, bu yüzden sevmiyorum.	Olumsuz
Matematikte başarılı olmak benim için önemli, bu yüzden çok çalışıyorum.	Tarafsız
Matematik dersinde kendimi yetersiz hissediyorum ve bu beni üzüyor.	Olumsuz
Matematik öğretmenimiz konuları çok iyi anlatıyor, bu yüzden matematik dersini seviyorum.	Olumlu
Matematik dersinde zorlandığım için sık sık stres oluyorum.	Olumsuz
Matematik, gelecekteki kariyerim için önemli olduğundan, bu dersi dikkatle takip ediyorum.	Tarafsız
Matematik problemlerini çözdükçe kendime güvenim artıyor.	Olumlu
Matematik dersleri bana çok karmaşık geliyor ve bu da motivasyonumu düşürüyor.	Olumsuz
Matematikte yeni şeyler öğrenmek beni heyecanlandırıyor.	Olumlu

```
If num_positive ≥ num_negative:
    sentiment = "Olumlu"
elif num_negative > num_positive:
    sentiment = "Olumsuz"
else:
    sentiment = "Tarafsız"
print("Duygusal ton:", sentiment)
Duygusal ton: Tarafsız
```



Appendix 5.

Example of Item Generation

Below are examples of open-ended and multiple-choice questions generated using GPT-4 and Gemini.

Text: Küçük Prens ve Tilki

> Bir zamanlar küçük prens bir gezegende yaşar. Bu gezegenin herhangi bir yerinde, özel bir çölde, altı çalı çırpıyla kaplı bir yere gömülmüş bir yıldız var. Bir gün küçük prens, çölde bir tilki ile karşılaşır. Tilki, küçük prensin onu evcilleştirmesini ister. Küçük prens, tilkinin ne demek istediğini anlamaya çalışır. Tilki, "Evcilleştirme, özlemin derecesine bağlıdır. Benimle ilgilenirsen, benimle dost olursan, benimle oyun oynarsan, senin için çok farklı bir ışık oluşur. Yıldızları seyretmek güzel olur. İnsanlar, bir yıldızı seyrettiğinde, senin yıldızında olduğun saatte gülümseyeceklerdir." der. Küçük prens tilkiyi evcilleştirmeyi kabul eder. Tilki, küçük prense insanların ne anlama geldiğini anlatır. Ona göre insanlar, birbirinden farklı olan

gülüşlerdir. Tilki, küçük prense insanların onları evcilleştirenlerdir. Evcilleştirmek ise, birbirine alıştırmaktır. Küçük prens, tilkinin anlattıklarını düşünür ve onunla dost olur.

Appendix Table 3.
Open-Ended Questions Generated with GPT-4o

Question Number	Question
1	Tilki, Küçük Prens'e "evcilleştirilmek" terimini nasıl açıkladı?
2	Küçük Prens ve tilki neden her gün aynı saatte buluştu?
3	Tilki, Küçük Prens'in gezegenine geri döneceğini öğrendiğinde nasıl hissetti ve neden?
4	Küçük Prens, tilkiyi evcilleştirmenin sonunda hangi önemli dersi öğrendi?
5	Tilki'nin, "İnsan ancak yüreğiyle baktığında doğruyu görebilir. Gözler gerçeği göremez." sözü ne anlama gelir?

Appendix Table 4.
Examples of Multiple-Choice Questions Generated with GPT-4o

Question	Options	Correct Answer
KüçükPrens Dünya'da kiminle karşılaştı?	a) Tilki b) Yılan c) Kral d) İş adamı	a) Tilki
Tilki, Küçük Prens'e ne olmayı teklif etti?	a) Düşman b) Öğretmen c) Arkadaş d) Rehber	c) Arkadaş
Küçük Prens tilkiyi evcilleştirmek için ne yaptı?	a) Onu besledi b) Onunla konuştu c) Onunla oyun oynadı d) Onun yanına her gün aynı saatte gitti	d) Onun yanına her gün aynı saatte gitti
Tilki, Küçük Prens'e neyi öğretti?	a) Zamanın değerini b) Arkadaşlığın önemini c) Cesaretin gücünü d) Bilginin değerini	b) Arkadaşlığın önemini
Tilki, Küçük Prens'e veda ederken ne dedi?	a) "Seni unutmayacağım." b) "Her zaman buğday tarlalarına bakacağım ve seni hatırlayacağım." c) "Geri dön." d) "Seni seviyorum."	b) "Her zaman buğday tarlalarına bakacağım ve seni hatırlayacağım."

Appendix Table 5.
Open-Ended Questions Generated with Gemini-1.5-Flash

Question
1. "İnsanlar, yaşadıkları dünyayı anlamak için sürekli yeni keşiflere ve öğrenmelere ihtiyaç duyarlar." cümlesini kendi cümlelerinizle açıklayınız. Bu cümle size ne düşündürüyor?
2. Bir yazarın romanında, kahramanın yaşadığı zorlukları ve bunlarla başa çıkma yollarını ele aldığını düşünün. Sizce yazar bu romanı neden yazmış olabilir? Bu soruyu yanıtlamak için romanın konusu ve kahramanın özellikleri hakkında tahminlerde bulunabilirsiniz.
3. "Herkesin bir hikayesi vardır." sözüyle ne anlatılmak isteniyor olabilir? Bu sözü destekleyen kendi yaşamınızdan bir örnek veriniz.
4. "Teknoloji geliştikçe, insanlar arasındaki iletişim biçimleri de değişiyor." cümlesini ele alarak, teknolojinin iletişim üzerindeki olumlu ve olumsuz etkilerinden bahsediniz.
5. Bir şiirde, yazarın doğayı, insanları veya duyguları nasıl anlattığını düşünüyorsunuz? Bir şiir örneği vererek, şiirde kullanılan dilin ve imgelerin okuyucunun duygularını nasıl etkilediğini açıklayınız.

Appendix Table 6.
Examples of Multiple-Choice Questions Generated with Gemini-1.5-Flash

Question	Response
"Ağaçlar, doğanın akciğerleridir." cümlesi aşağıdaki hangi anlamı taşır?	a) Ağaçlar, doğada önemli bir rol oynar b) Ağaçlar, insanlara oksijen sağlar. c) Ağaçlar, doğanın güzelliğini temsil eder. d) Ağaçlar, insanlara gölge yapar.
"Yazarın romanında, kahramanın yaşadığı zorluklar ve bunlarla başa çıkma yolları ele alınıyor." cümlesine göre, romanın ana teması nedir?	a) Aşk ve romantizm b) Savaş ve kahramanlık c) Yaşamın zorlukları ve üstesinden gelme d) Teknoloji ve gelecek
"İnsanlar, yaşadıkları dünyayı anlamak için sürekli yeni keşiflere ve öğrenmelere ihtiyaç duyarlar." cümlesinde vurgulanan ana fikir nedir?	a) İnsanlar meraklıdır. b) İnsanlar öğrenmeyi sever. c) İnsanlar bilginin önemini anlar. d) İnsanlar sürekli gelişen bir dünyada yaşar.
"Teknoloji geliştikçe, insanlar arasındaki iletişim biçimleri de değişiyor." cümlesinde, teknolojinin iletişim üzerindeki etkisi nasıl gösteriliyor?	a) Teknoloji, iletişimi daha hızlı ve kolay hale getiriyor. b) Teknoloji, iletişimi daha kişisel hale getiriyor. c) Teknoloji, iletişimi daha karmaşık hale getiriyor. d) Teknoloji, iletişimi daha az etkili hale getiriyor.
"Herkesin bir hikayesi vardır." sözüyle ne anlatılmak isteniyor olabilir?	a) Her insanın yaşamında önemli olaylar vardır. b) Her insanın benzersiz bir yaşam hikayesi vardır. c) Her insanın anlatmaya değer bir hikayesi vardır. d) Her insanın başkalarına anlatabileceği bir hikayesi vardır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Technical Brief; 2024; 15(4); 446-451

FAfA: Factor Analysis for All An R Package to Conduct Factor Analysis with R Shiny Application

Abdullah Faruk KILIÇ *

Abstract

This paper presents a presentation of FAfA (the R Shiny application), which was specifically developed for performing complete factor analysis processes. These procedures include data wrangling, assumption checks for exploratory and confirmatory factor analysis, reliability analysis, exploratory graphic analysis, and item weighting. The objective of the paper is to provide users with clear instructions on how to effectively use the FAfA package, therefore guaranteeing precise and consistent outcomes in their research. The FAfA application's primary goal is to integrate EFA and CFA into a single software. Furthermore, FAfA possesses the capability to compute several reliability coefficients related to internal consistency. It can also be utilized when item weighing is desired. This package is advantageous as it enables the verification of assumptions prior to analysis within a single program, facilitates both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) in one application, provides reliability coefficients not accessible in user-interface programs (such as stratified alpha), and integrates exploratory graph analysis, which has rapidly advanced in recent years, into a unified application.

Keywords: factor analysis, reliability analysis, item weighting, exploratory graph analysis

Introduction

Factor analysis (FA) is a widely used method to collect validity evidence for measures. There are numerous software tools available for conducting factor analysis (FA). SPSS can conduct exploratory factor analysis (EFA), Mplus and AMOS can conduct confirmatory factor analysis (CFA), and Factor software is used for EFA. However, there are limitations to the software. First, no module in the software allows for a stand-alone examination of EFA or CFA assumptions. Furthermore, it is typically unfeasible to conduct both EFA and CFA using a single software. While JASP or JAMOVI provide a solution to this challenge, it is essential to note that verifying the assumptions in these software platforms also requires a significant amount of time. In contrast, the FAfA R Shiny application is designed to save researchers time, allowing them to focus on their analysis and interpretation. Furthermore, when conducting EFA in SPSS, the Pearson correlation matrix is utilized. However, due to the increased collection of ordinal data in domains like education and psychology, it may be necessary to use a polychoric correlation matrix. In addition, removing outliers in this software (JAMOVI and JASP) is complicated. So, I created an RShiny app named FAfA to conduct EFA, CFA, and reliability analysis with data wrangling and assumption check properties.

Dependencies of FAfA Application

In any Shiny application, it is essential to ensure that all necessary packages are loaded. FAfA uses shiny (Chang et al., 2024) for building the user interface, *dplyr* (Wickham et al., 2023) for data manipulation, *psych* for EFA (Revelle, 2024), *lavaan* (Rosseel, 2012) for CFA, *EGAnet* (Golino & Alexander, 2023) for exploratory graph analysis. I used the *psych* (Revelle, 2024) for the alpha coefficient, *MBESS* (Kelley, 2023) for omega, *semTools* (Jorgensen et al., 2022) for structure reliability, *sirt* (Robitzsch, 2021) for stratified alpha, and I wrote code for Armor's theta reliability coefficient. Loading these packages at the beginning ensures that all dependencies are available when needed, which is a best practice in R programming.

* Associate Professor, Trakya University, Faculty of Education, Edirne-Türkiye, afarukkilic@trakya.edu.tr, ORCID ID: 0000-0003-3129-1763

To cite this article:
Kılıç, A.F. (2024). FAfA: Factor analysis for all an R package to conduct factor analysis with R Shiny application. *Journal of Measurement and Evaluation in Education and Psychology, 15*(4), 446-451. https://doi.org/10.21031/epod.1555805
Received: 25.09.2024
Accepted: 17.12.2024

Reading User-Uploaded Data

FAfA analyzes the data formatted by ".dat," ".txt" (Text Document—MS-DOS Format), or ".prn" file. Variable names should not be included in the data set, and missing data should be indicated with NA. The application ensures that the uploaded data is read correctly and is ready for subsequent processing steps. Figure 1 illustrates the data example and the application's control result.

Figure 1

Data Set and Control Results of Data

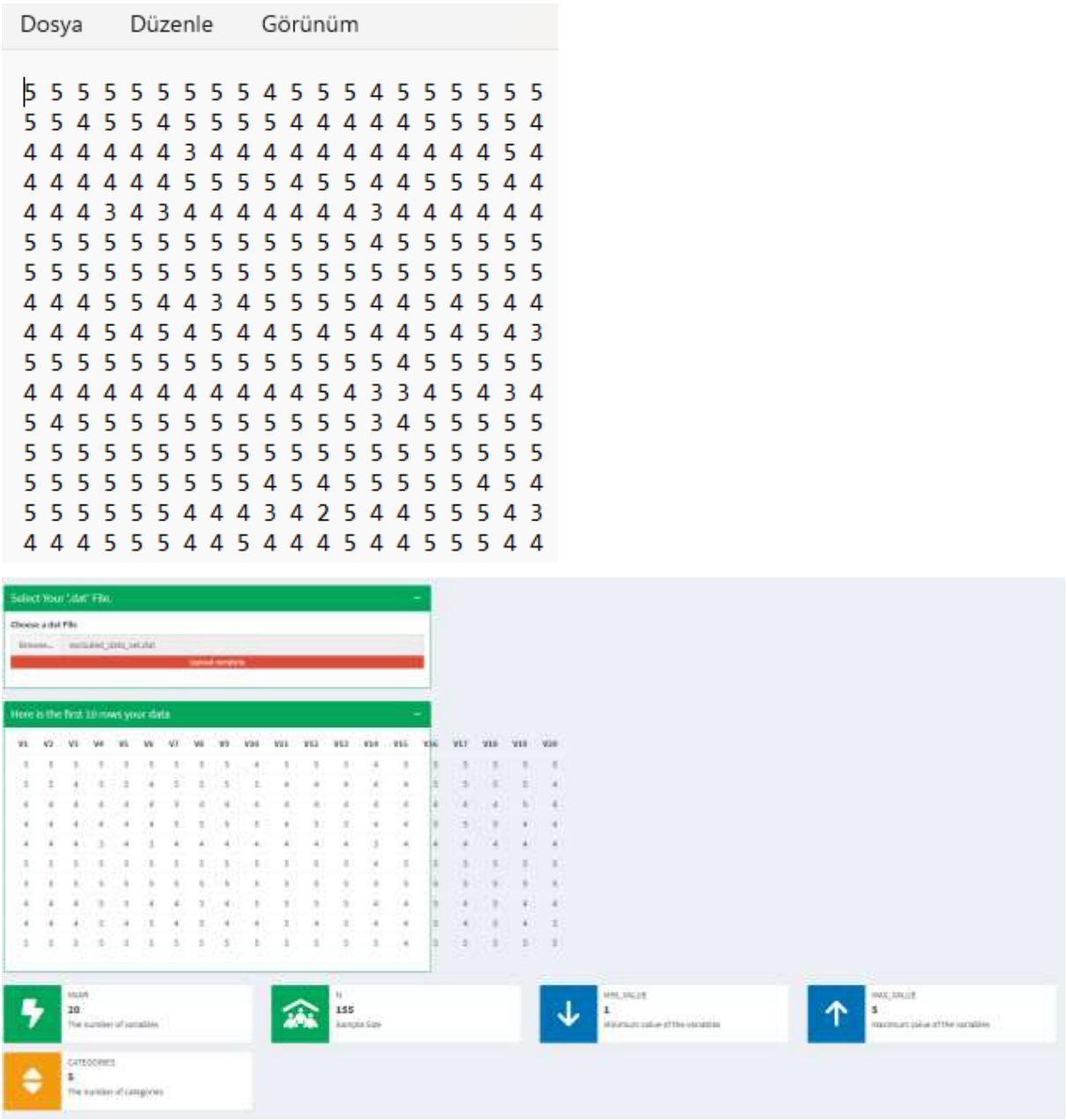


Figure 1 displays the view of the data set and the results of FAfA's check of the data set, which was performed after reading the data in the FAfA application. The dataset has a ".dat" extension, and there are no variable names (column names). After reading the dataset, the first 10 rows are identified as FAfA output. Then, the number of variables in the dataset, sample size, minimum value and maximum value in the dataset, and number of categories are displayed. Thus, it can be examined whether the data set is read correctly or not.

Wrangling Data

In order to exclude the variables, the column numbers of the variables to be excluded should be written by placing a comma between them. Once you have defined the variables in this manner, use the "Exclude the variables button" to exclude them. Next, save the data set containing the excluded variables to your computer using the download excluded data button, and then read it again in the FAfA application. FAfA can randomly split the data set into two. To accomplish this, first split the data set into two using the "split my data" button, then "download the EFA data" to save the first half on your computer, and download the CFA data to save the second half. Similarly, whatever data set is analyzed should be read again in the FAfA application. This should not be interpreted as applicable to every data set. In scale development research, data is collected again following the examination of the scale's structure using EFA, after which Confirmatory Factor Analysis (CFA) is conducted. Utilize this part to circumvent conducting EFA and CFA on a singular data set, as indicated by Fokkema and Greiff (2017). FAfA uses the Mahalanobis distance statistic to identify multivariate outliers. Accordingly, those that are significant at the a=0.001 level are examined in the Examined Outliers section. This section reports how many outliers there are in the data set. Additionally, you can remove outliers from the data set by using the "Remove outliers from my data" button. Next, save the data set without outliers to your computer using the "Download the data set without outliers" button, and then read it back into FAfA.

Excluding Variables

The FAfA application incorporates the capability to eliminate user-specified variables from the dataset. Users can designate specific variables to exclude. Excluding variables is often essential for a variety of reasons, including eliminating irrelevant variables, correcting multicollinearity, managing missing data, or reducing item dropouts, which are consequences of EFA. This function also aids in data preparation for more sophisticated studies by eliminating potential sources of bias or noise.

The method of removing variables entails enabling users to input indices (column numbers) to be eliminated from the dataset. Subsequently, the application generates a new dataset by removing the given variables. The new dataset must be uploaded to the application again.

Assumptions

In the provided code, the assumption function performs several statistical tests and calculations to check the assumptions required for EFA and CFA. This function is a crucial component of the application, as it ensures that the data meets the necessary criteria for valid statistical analysis. The function finds outliers, checks for multicollinearity, and checks for multivariate normality using Mardia's multivariate skewness and kurtosis values (Mardia, 1970). It also finds out what the minimum and maximum values are and the number of missing data points.

The application generates descriptive statistics and checks assumptions for further analysis. This includes calculating various descriptive measures, checking for multicollinearity, and assessing normality. The results are displayed in the application, and the user can download them. Providing descriptive statistics and assumption checks helps users gain a comprehensive understanding of their data. It allows them to identify potential issues and make informed decisions about the appropriate analysis techniques to use. By incorporating these functionalities into the FAfA, users can perform thorough and reliable data analyses.

Exploratory Graph Analysis (EGA)

Exploratory Graph Analysis (EGA) is a technique used to identify the data's underlying structure by estimating the network structure. EGA is a novel technique introduced in the field of network psychometrics to determine the number of factors that underlie multivariate data. EGA generates a network plot that provides a visual representation of the optimal number of dimensions to keep. Additionally, this plot reveals the clustering of items and the strength of their associations (Golino et al., 2020). The FAfA application performs EGA and provides a visualization of the network, helping users identify clusters or groups of variables that are correlated. Figure 2 demonstrates the EGA results of an example data set.

Figure 2
EGA Results of the FAfA Package

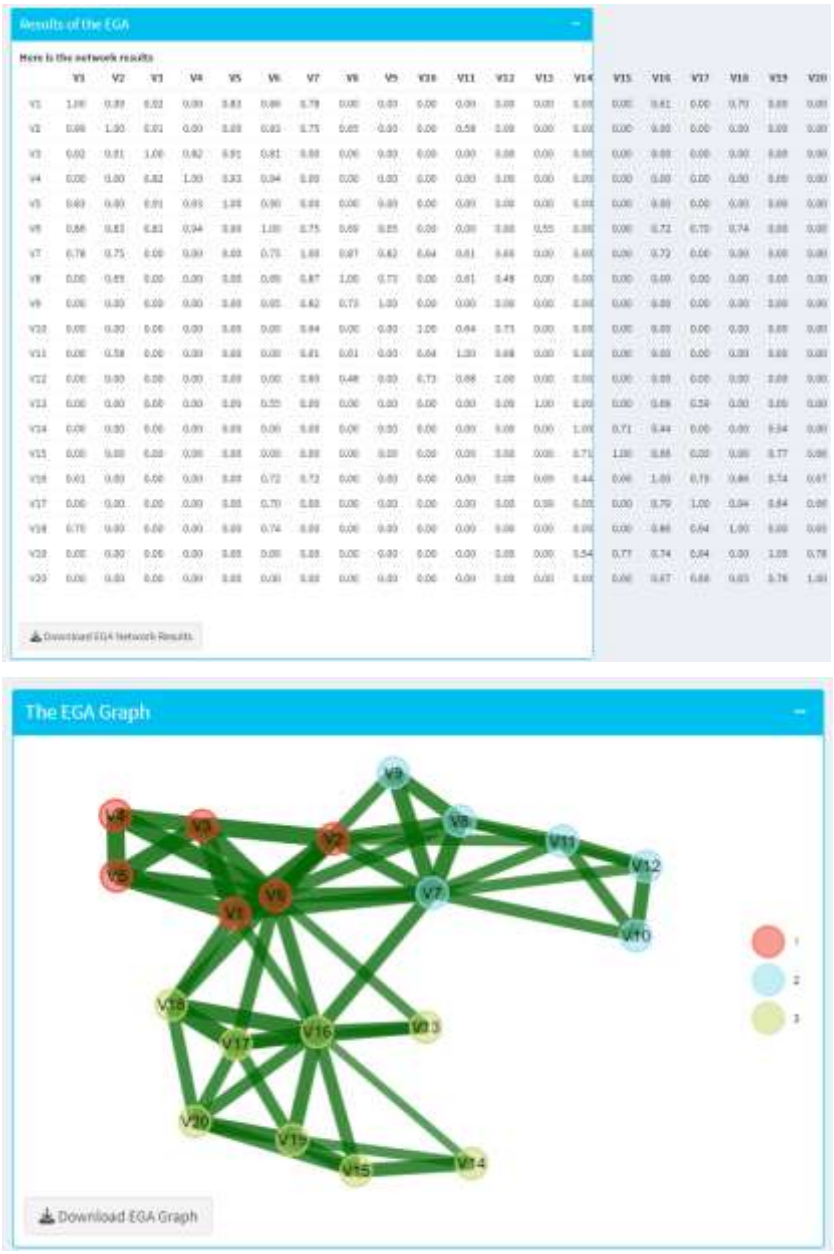


Figure 2 illustrates that FAfA initially presents the outcomes derived from the EGA analysis, followed by the corresponding network graph. The example demonstrates a three-factor structure and strong correlations among the items.

Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis (EFA) is a statistical technique used to identify the underlying factor structure of the data (Gorsuch, 1974). The FAfA application is utilized for EFA and employs multiple techniques, such as parallel analysis (Horn, 1965) and the Hull approach (Lorenzo-Seva et al., 2011), to ascertain the number of factors. Furthermore, the KMO statistic and the findings of Bartlett's test of sphericity, which are commonly assessed for EFA. FAfA also computes the MSA index, as proposed by Lorenzo-Seva and Ferrando (2021), for each item. Subsequently, FAfA will depict the correlation matrix among the items utilizing a heat map. The results of the factor analysis, including factor loadings and explained variance, are displayed and can be downloaded by the user. Users can conduct the EFA with Pearson or a polychoric/tetrachoric correlation matrix. Various extraction methods, such as

principal axis factoring or maximum likelihood estimation, can be used. Oblique and orthogonal rotation methods, such as varimax or oblimin, help achieve a more straightforward factor structure by maximizing the variance explained by each factor. The FAfA provides options for users to choose the extraction and rotation methods, ensuring flexibility in the analysis.

Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis (CFA) validates the factor structure known prior to the analysis (Brown, 2015. Users can define their factor structure, and the FAfA estimates the model and provides various fit measures to assess the model fit. The results, including factor loadings and modification indices, are displayed in the FAfA output. Users can download these outputs. The CFA process entails specifying a factor model in which the relationships between observed variables and latent factors are defined. The model is then estimated using techniques such as maximum likelihood estimation. FAfA reports fit measures such as Chi Square, degrees of freedom (df), Chi Square/df, p value of Chi Square, CFI, TLI, RFI, SRMR, RMSEA and its 90% confidence interval. They are used to assess the model's fit to the data. The application provides a summary of these fit measures, helping users evaluate the adequacy of their specified factor model. In addition, model modification suggestions can be examined.

Reliability Analysis

Reliability analysis assesses the internal consistency of the data, providing a measure of the reliability of the scales used (Mueller & Knapp, 2019). The FAfA calculates various reliability coefficients, such as Cronbach's alpha and McDonald's omega. The results are displayed, and the user can download them. Assessing reliability is essential in psychometrics. FAfA calculates not only the unidimensional reliability coefficients, but also a multidimensional reliability coefficient named stratified alpha. Stratified alpha has been proposed to calculate the reliability of composite scores obtained from measurement tools with multiple subdimensions (Cronbach et al., 1965).

Item Weighting

I added the item weighting function to enhance the validity of the measures suggested by Kılıç and Doğan (2019). Item weighting is a valuable technique for refining measurement scales. By adjusting scores based on item properties, users can enhance the construct validity and reliability of their scales. The FAfA provides a convenient tool for implementing item weighting, helping users improve their measurement instruments. This weighting function assigns a weight to the item based on the combined values of item difficulty and the respondent's average score. If this sum exceeds 1, the item reliability is incorporated into the respondent's answers. If this sum does not exceed 1, the respondent's score remains unchanged (1 for a correct item, 0 for an incorrect item).

Conclusion

This manuscript provides a detailed review of an R Shiny application named FAfA. This comprehensive review serves as a guide for researchers and practitioners looking to utilize FAfA for their data analysis, enhancing the accessibility and usability of advanced statistical techniques. The R package FAfA is available on the Comprehensive R Archive Network (CRAN; http://www.cran.r-project.org). FAfA can be installed with `install.packages("FAfA", dependency = TRUE)` code in R. The FAfA package requires additional packages. The FAfA package specifies these packages in its suggestions section. If you get an error after running that installation code, running the following code may also help you prevent errors.

```
install.packages(c("EFA.MRFA", "EFA.dimensions", "EFAtools", "EGAnet", "MBESS", "config", "dplyr", "energy", "ggcorrplot", "golem", "lavaan", "mctest", "moments", "mvnormalTest", "pastecs", "psych", "psychometric", "semPlot", "semTools", "shiny", "shinycssloaders", "shinydashboard", "sirt"))
```

After loading the packages with these scripts, we may invoke the package using the `install.packages(FafA)`. A function exists within the package. This function is the `run_app()` function. The `run_app()` function executes the package directly in English. The function

`run_app(lang = "tr")` can be utilized to activate Turkish language support if preferred. Source code and documentation are freely available from https://CRAN.R-project.org/package=FAfA.

Declarations

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

Ethical Approval: I confirm that I have followed all ethical guidelines for authorship. This study does not necessitate ethical approval due to its nature as a software presentation.

References

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R.* (Version 1.7.5) [R package]. https://CRAN.R-project.org/package=shiny

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement, 25*(2), 291–312. https://doi.org/10.1177/001316446502500201

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble. *European Journal of Psychological Assessment, 33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Golino, H., & Alexander, C. P. (2023). *EGAnet: Exploratory Graph Analysis—A framework for estimating the number of dimension in multivariate data using network psychometrics.* (Version 2.1.0) [R package]. https://cran.r-project.org/web/packages/EGAnet/EGAnet.pdf

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., Thiyagarajan, J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods, 25*(3), 292–320. https://doi.org/10.1037/met0000255

Gorsuch, R. L. (1974). *Factor analysis.* W. B. Saunders.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (Version 0.5-6) [R package]. https://CRAN.R-project.org/package=semTools

Kelley, K. (2023). *MBESS.* (Version 4.9.3) [R package]. https://cran.r-project.org/package=MBESS

Kılıç, A. F., & Doğan, N. (2019). The Effect of item weighting on reliability and validity. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 10*(2), 148–163. https://doi.org/10.21031/epod.516057

Lorenzo-Seva, U., & Ferrando, P. J. (2021). MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis. *Methodology, 17*(4), 296–306. https://doi.org/10.5964/meth.7185

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

Mueller, R. O., & Knapp, T. R. (2019). Reliability and validity. In G. R. Hancock, L. M. Stapleton, & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (2nd. ed., pp. 397–401). Routledge.

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.4.3) [R package]. https://cran.r-project.org/package=psych

Robitzsch, A. (2021). *sirt: Supplementary Item Response Theory Models.* (Version 3.9-3) [R package]. https://cran.r-project.org/package=sirt

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

Wickham, H. (2021). *Mastering shiny*. O'Reilly Media.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of data manipulation.* (Version 1.1.3) [R package]. https://CRAN.R-project.org/package=dplyr