Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Güz* 2022 *Cilt: 13- Sayı: 3*
Autumn 2022 *Volume: 13- Issue: 3*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

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.

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

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

Halil YURDUGÜL (Hacettepe Üni.)
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.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü 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)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serkan ARIKAN (Boğaziçi Üni.)
Seval KIZILDAĞ ŞAHİN (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Süleyman DEMİR (Sakarya Üni.)
Sümeyra SOYSAL (Necmettin Erbakan Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT (İzmir Demokrasi Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Wenchao MA (University of Alabama)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Yusuf KARA (Southern Methodist University)
Zekeriya NARTGÜN (Bolu Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni.)

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

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 164-193

A Bibliometric Analysis: A Tutorial for the Bibliometrix Package in R Using IRT Literature

Serap BÜYÜKKIDIK*

Abstract

The bibliometrix package in R programming language, which is frequently used in bibliometric analysis, was introduced in this research. The article aimed to illustrate the various analyses applied in a bibliometric study. For this purpose, articles containing the "item response theory" (IRT) or "item response modeling" or "item response model" terms in the abstract were searched in the Thomson Reuters Clarivate Analytics Web of Science (WoS at http://www.webofknowledge.com), and bibliometric data was downloaded. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) steps were followed in the study. Data from 3388 IRT-related articles on education and psychology, searched between 2001 and 2021, were used in the study. Data were analyzed with the *bibliometrix* package. Some of the stages in data analysis were shared with screenshots. As a result of data analysis through the real data set, the author's keywords related to IRT were item response model, differential item functioning, psychometrics, assessment, measurement, reliability, validity, Rasch model, and measurement invariance. The countries with the highest number of citations in IRT studies were the USA, Canada, Netherlands, United Kingdom, and China, respectively. Turkey ranked 12th in IRT studies with 434 citations. It was thought that bibliometric analysis of articles related to IRT would shed light on researchers in the field of psychometrics.

Keywords: bibliometric analysis, item response theory (IRT), biblioshiny, bibliometrix, R

Introduction

At the heart of science is the desire to know. Heike Kamerlingh Onnes (1882) said "Measuring is knowing" (as cited in van Raan, 2004, p. 21). Lord Kelvin said, "One's knowledge of science begins when he can measure what he is speaking about, and express it in numbers." (as cited in Eysenck, 1973). Horace (65–5 BC) said, "There is a measure in all things (Est modus in rebus)" (as cited in van Raan, 2004, p. 21). The scientific information produced today is increasing compared to the past (Linnenluecke et al., 2020; Ware & Mabe, 2015). In the globalizing world, borders are no longer important, information is spreading rapidly, and scientific knowledge is increasing. Science and technology change rapidly in the information age, and countries that follow this change achieve economic growth and social welfare and keep up with the competitive world as independent countries (National Science Foundation (NSF), 2007, p. vii). Two of the development indicators of science and technology are "education and advanced training" and "scientific publications, collaboration, and citations" (NSF, 2007, p. viii). Analysis of scientific publications, collaborations, and citations is possible with bibliometric research. Bibliometrics, scientometrics, informetrics, and librametrics were similar but non-synonymous concepts consisting of the combination of bibliography, science, knowledge, and the library, and the word "metrics", respectively (Egghe, 2005, p. 1311; Sengupta, 1992, p. 25). All these concepts were directly related to the measurement of information. Bibliometry was used in the construction of knowledge and the development of new ideas (Sengupta, 1992, p. 25).

The word bibliometric was derived from the Greek and Latin word "biblio", which means book, and the word "metrics," which refers to the measurement (Sengupta, 1992, p. 25). The term statistical bibliography was first used by E. Wyndham Hulme in 1922 as part of a course at the University of

* Assistant. Prof., Sinop University, Faculty of Education, Sinop-Turkey, sbuyukkidik@gmail.com, ORCID ID: 0000-0003-4335-2949

To cite this article:
Büyükkıdık, S. (2022). A bibliometric analysis: A tutorial for the bibliometrix package in R using IRT literature. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 164-193. https://doi.org/10.21031/epod.1069307
Received: 7.02.2022
Accepted: 1.09.2022

Cambridge (Pritchard, 1969). However, bibliometric studies had their origins dating back to the 1890s (Sengupta, 1992, p. 25). The main breakthrough in the improvement of bibliometrics was Garfield's (1955, 1964) development of a Science Citation Index (bibliographic databases). Bibliometric data began to be recorded with Science Citation Index. The purpose of bibliometrics is basically to evaluate the scientific literature in the relevant field. Therefore, researchers can apply bibliometrics to any field of science (Andrés, 2009, p. 1).

Due to the rapid publication of scientific research and the fact that there are many journals, the quality of scientific publications produced may decrease (Demir, 2018). Many bibliometric data such as citation numbers, keywords, titles, collaborations, and institutions can be produced from various databases. Bibliometrics research can be carried out using these databases. Bibliometric research can be discussed in terms of characteristics of the publications, citation impact, country analysis, and subject analysis. The bibliometric analysis differs from the systematic review and meta-analysis (Donthu et al., 2021).

Bibliometric analysis techniques were divided into main techniques and enrichment techniques. Main techniques include "performance analysis (publication-related metrics, citation-related metrics, citation-and-publication-related metrics) and science mapping (citation analysis, co-citation analysis, bibliometric coupling, co-word analysis, and co-authorship analysis)". Enrichment techniques include network analysis (network, metric, clustering, visualization) (Donthu et al., p. 288).

Bibliometric analysis methods

Bibliometric analysis methods are classified in various ways (e.g., Aria & Cuccurullo, 2017; Donthu et al., 2021; Durieux & Gevenois, 2010; Zupic & Cater, 2015). This study discussed citation analysis, co-citation analysis, co-author analysis, co-word, and bibliometric coupling analysis.

Co-author

In this method, links are established between co-authors of an article. The unit of analysis is the authors. This analysis reveals the networks and collaborations between the authors and their countries and institutions. The downside is that co-authoring is not always indicative of collaboration (Zupic & Cater, 2015).

Co-word

A co-word analysis is done by considering the keywords, titles, and common words in the abstract. The unit of analysis is words. Most importantly, it uses the actual content of the documents for analysis, while other methods use only bibliographic metadata. In addition, this method does not take into account the negative feature of words being handled in different ways and having different meanings (Zupic & Cater, 2015).

Citation analysis

Citation analysis is used to analyze the effectiveness of authors, documents, or journals by considering their citation rates. The unit of analysis is documents, authors, and journals. Its positive features are effective in finding essential studies in the field. The negative feature is that the cumulative citation rate of new publications will be low, the most citations are likely to be to older publications (Zupic & Cater, 2015).

Co-citation

This analysis analyzes the links between authors, documents, or journals by considering reference lists. As in citation analysis, the unit of analysis is authors, documents, and journals. Pros are that it is the most valid and frequently used bibliometric method (Zupic & Cater, 2015).

Bibliometric coupling

In this method, links are established between documents, authors, and journals based on references. The analysis unit is journals, documents, and authors, as in citation and co-citation methods. The positive aspect of this method is that it can be used immediately without the need to accumulate citations. It is crucial to reveal new publications and fields that have not been cited. The disadvantages of this method are that it can only be used for a limited time. Since this method does not focus on citations, it is difficult to know whether mapped publications are important (Zupic & Cater, 2015). Detailed information about which bibliometric method answers which research questions can be found in the literature (see Zupic & Cater, 2015).

In this research, articles related to Item Response Theory (IRT) were investigated. There were various theories in the emergence of psychometric characteristics in education and psychology. Two of these theories were Classical Test Theory (CTT) and IRT. The first signs of IRT can be seen in Louis Leon Thurstone's (1925) article entitled "A Method of Scaling Psychological and Educational Tests" (Bock, 1997, p. 21). The onset of IRT was usually based on Lord and Novick's (1968) classic textbook, “Statistical Theories of Mental Test Scores” in the United States (Embretson & Reise, 2013, p. 5). In recent years, research on IRT has increased due to the superiority of IRT over CTT. For the last 30 years, IRT has been frequently used by the largest test companies in the world for "design of tests, test assembly, test scaling, and calibration, construction of test item banks, investigations of test item bias and other common procedures in the test development process" (Hambleton et al., 1991, p. VII). Embretson and Reise (2013, p. 249) discussed IRT applications under three headings as differential item functioning (DIF), computerized adaptive testing (CAT), and scale construction. Besides this, IRT is used in many areas of psychometrics, such as test equating and standard-setting. In this bibliometric research, articles in WoS related to IRT were handled. After the 2000s, IRT research in WoS has shown a great improvement. In order to address this development, articles related to IRT after 2001 were taken into consideration in this study.

The Importance of Bibliometric Research

Synthesizing past research findings and bibliometric data is one of the most important steps for the cumulative advancement of scientific knowledge. Bibliometric research makes an objective assessment of the scientific literature using a quantitative approach. These studies provide a transparent, systematic, and reproducible literature review. This bibliometric review sheds light on those who want to research the relevant literature (Pritchard, 1969, p. 348; Zupic & Cater, 2015, p. 1). Bibliometrics is a set of statistical and mathematical methods used to measure and analyze the quality and quantity of articles, books, and other publications (Durieux & Gevenois, 2010, p. 342). Information and communication activities and scientific documentation are developed through a quantitative analysis of the production, dissemination, and use of information obtained from library collections and services such as WoS in bibliometric research. Thus, it is aimed to contribute to a better understanding of the scientific research mechanism in bibliometric research (Osareh, 1996, p. 150).

With bibliometric analysis, we gain information about both the intellectual structure and the conceptual framework. We get ideas about the progress of research on specific topics. In addition, bibliometric research helps journal editors make their decisions and evaluations (Zupic & Cater, 2015, p. 9). Durieux and Gevenois (2010, p. 342) mentioned three types of bibliometric indicators and their importance. These indicators are “quality indicators” that measure the quality (or "performance") of a particular

researcher's output; quantity indicators measuring the researcher's productivity; and structural indicators that measure the links between publications, authors, and research areas. Bibliometric indicators are vital for organizations and researchers in funding decisions, assignments, and promotions. Today, as more scientific discoveries occur, knowledge accumulation increases and bibliometric indicators are becoming more critical day by day (Durieux & Gevenois, 2010, p. 342).

When the bibliometric studies in the literature were examined, researches were found in the field of education (e.g., Gülmez et al., 2020), and in the field of educational administration (e.g., Gümüş et al., 2019; Hallinger & Hammad, 2019). In the field of measurement and evaluation in education, researches were conducted in the fields of CAT (e.g., Yurtçu & Güzeller, 2021), IRT (e.g., Aksu & Güzeller, 2019), and DIF (e.g., Gómez Benito et al., 2005). A completely similar study has not been found that performs bibliometric analysis of articles related to IRT according to specific criteria in WoS with an R tutorial.

It was aimed to conduct a bibliometric analysis of IRT articles searched in WoS with the *biblioshiny* interface obtained using the R *bibliometrix* package in this research. Another purpose of the research was to share the plots or figures, and tables obtained within the scope of the research by following the process steps in biblioshiny. In recent years, the number of bibliometric research has been increasing. Introducing the user-friendly *biblioshiny* interface was thought to guide future research.

Methods

Database, terms, inclusion and exclusion criteria

Research data were obtained from the Web of Science. The terms used in the study were "item response theory", "item response modeling" or "item response model". The terms "IRT" was also used in the research. However, since the abbreviation of the irrelevant words like "infrared thermography" was "IRT", this term was removed when the researches were examined. Initially, only the term item response theory was used, while the words "item response modeling" and "item response model" were also found at the end of the literature review. That's why these terms were included in the research. The inclusion criteria of the study (1) include "item response theory" or "item response modeling" or "item response model" terms in the abstract, (2) the publication type was the article, (3) the publications were published in the field of psychology, and education, (4) the publications were published between 2001 and 2021, and (5) all publications were searched articles in Social Sciences Citation Index (SSCI) or Emerging Sources Citation Index (ESCI) or Science Citation Index Expanded (SCI-E) or Arts and Humanities Citation Index (A&HCI).

Analysis of Data

Bibliometric analysis can be done with many software or package programs like CoPalRed (Bailón-Moreno et al., 2005), CitNetExplorer (van Eck & Waltman, 2014), SciMAT (Cobo et al., 2012), Bibexcel (Persson et al., 2009), BiblioMaps (Grauwin & Sperano, 2018), Sci^{2}Tool (Sci^{2}Team, 2009), Biblioshiny (Aria & Cuccurullo, 2017), CiteSpace (Chen, 2006), VOSviewer (van Eck & Waltman, 2010). Only two of them offer a web-based user interface. *Biblioshiny* was a free web-based interface with the R (R Core Team, 2021) operative system. R was a free open-source software. The steps to run the Biblioshiny interface were explained in order. To use the *"bibliometrix"* package (Aria & Cuccurullo, 2017), R programming language (R Core Team, 2021) must be downloaded first. To download the free up-to-date R programming language and information about the R programming language can be accessed from the cran website. You can have information about the R programming language by accessing the R introductory document prepared by Venables, Smith, and the R Development Core Team (2021). After downloading the R programming language, R Studio must be downloaded. The free version of R Studio can be accessed from the web address. After R Studio was installed, the *install.packages("bibliometrix")* command was run. At these stages, your computer must be connected

to the internet. Then the *bibliometrix* package was activated. The *library ("bibliometrix")* command was used for this purpose. The *biblioshiny()* command was also typed and executed to open the user-friendly interface. The image of R studio is given in Figure 1. After running *the biblioshiny()* command, the user-friendly *biblioshiny* interface opened (see Figure 2).

Figure 1

R Studio User Interface

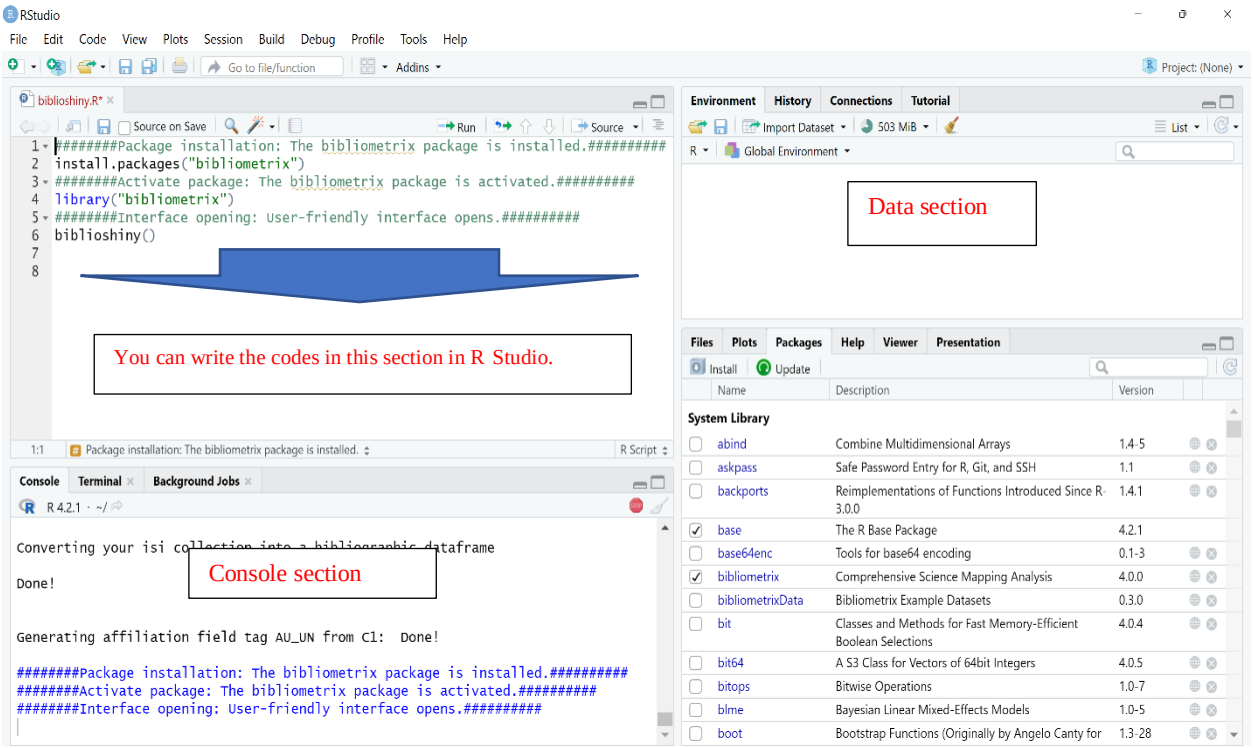
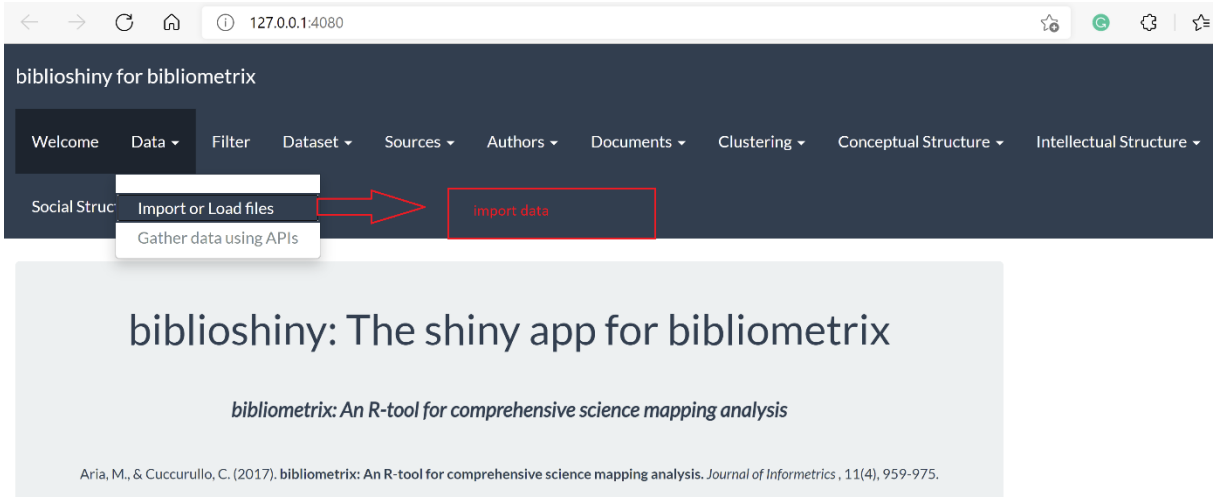


Figure 1 provides sections in R studio. Codes were written in the upper-right window and run in an orderly in Figure 1. As can be seen in Figure 2, this interface has 12 tabs. Bibliometric analysis can be performed by clicking on the tabs of the interface.

Figure 2

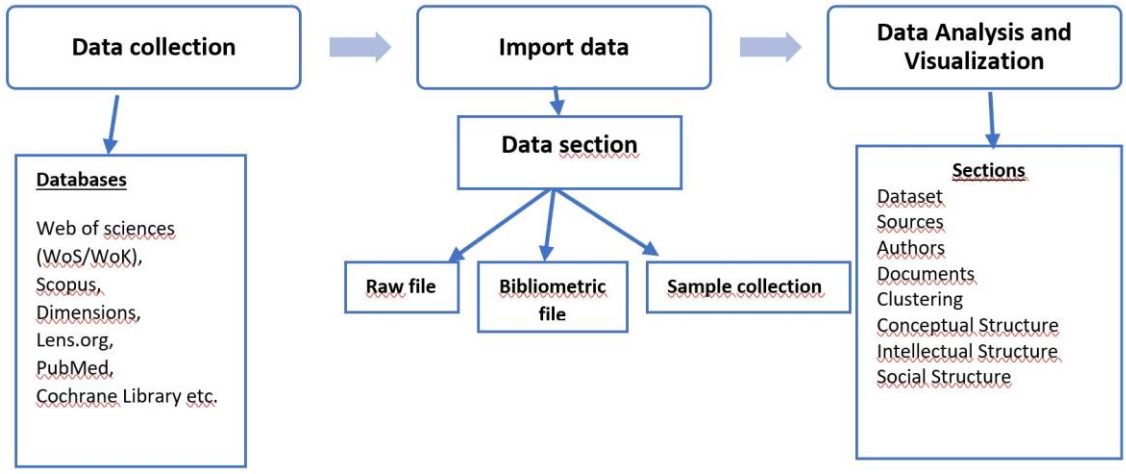
The Graphical User Interface (GUI) of the Biblioshiny and Loading the Data



It is seen that there were "data", "filter", "dataset", "sources", "authors", "documents", "clustering", "conceptual structure", "intellectual structure", "social structure", and finally "quit" options in Figure 2. Figure 3 displays bibliometric research's steps using *biblioshiny* (the shiny app for bibliometrix).

Figure 3

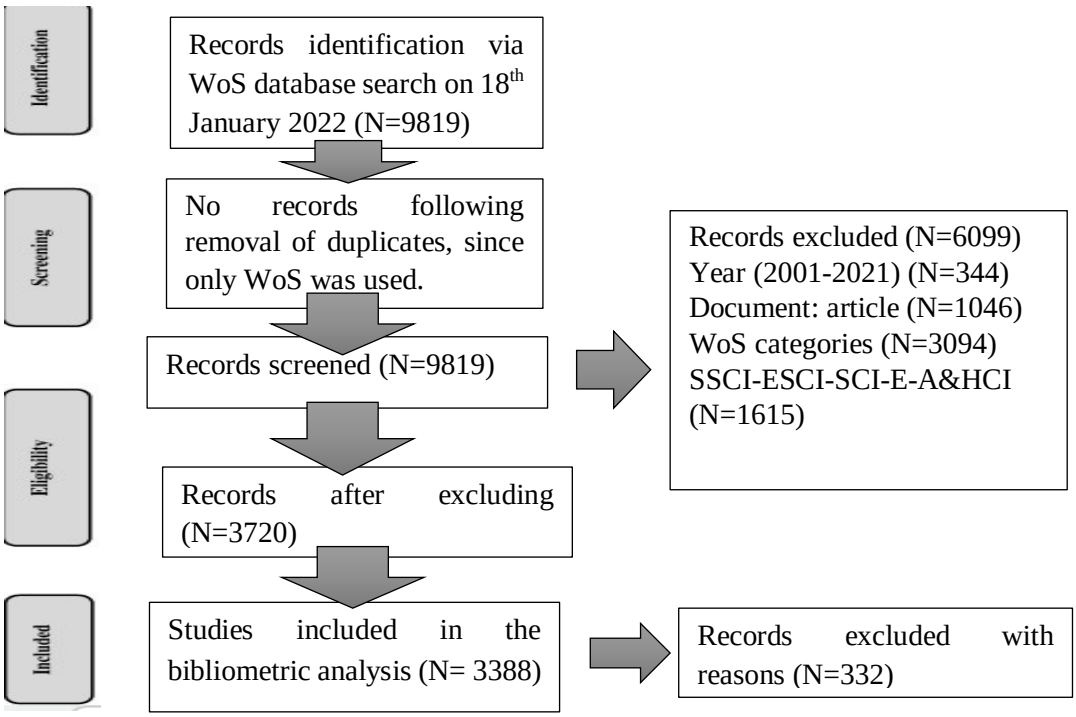
Bibliometric Research's Steps Using Biblioshiny



Bibliometric research consisted mainly of (1) decision on the research problem, (2) reviewing purpose-based literature (3) determining database, terms, inclusion and exclusion criteria, (4) decision on bibliometric method, and software to analyze data, (5) collection, and regulation of data within the framework of the research problem, and bibliometric method, (6) loading data into software and analyzing data, (7) visualizing, reporting findings, and writing impacts and recommendations. To analyze the data, bibliometric data must be uploaded to the *biblioshiny* first. Databases such as Web of Sciences (WoS/WoK), Scopus, Dimensions, Lens.org, PubMed, Cochrane Library can be used. WoS data was used within the scope of this research. PRISMA steps were followed in the study concerning the inclusion criteria of the research (see Moher et al., 2009). The PRISMA flow diagram used in the research is given in Figure 4.

Figure 4

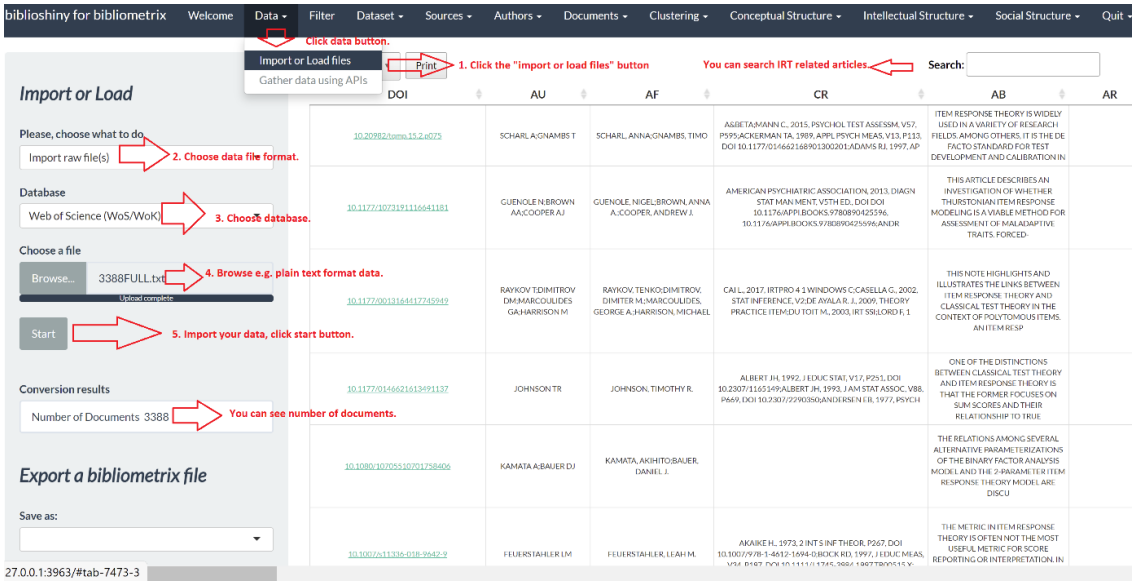
PRISMA Flow Diagram Steps in the Bibliometric Research



The data set obtained by following the PRISMA flow diagram steps had been downloaded from WoS in plain text format. Some records were extracted from plain text based on exclusion criteria, and the final file was created. In the study, analyses were carried out on 3388 publications related to IRT. Figure 5 shows a screenshot of uploading data in the *biblioshiny* interface.

Figure 5

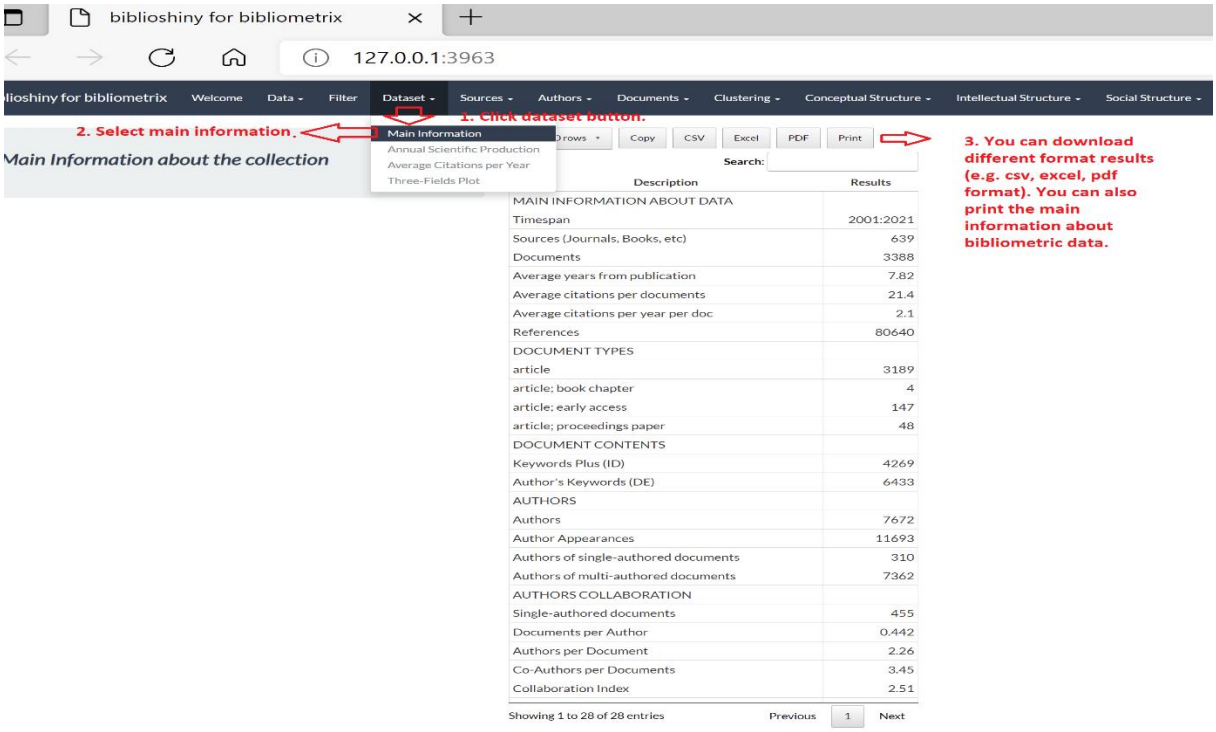
The View of the "3388" Dataset



In the interface that opened in the first step to analyze data downloaded from WoS, the import or load files option was clicked under the data tab. Then the data format was selected. This format can be a raw data file, a bibliometric file, or a sample collection. *"Import raw data file(s)"* was selected for the "*please choose what to do"* section because we used WoS data in the research. In the third step, "*Web of Science (WoS/WoK)"* was selected as the *database*. Data from databases such as Scopus, Dimensions, Lens.org, PubMed, and Cochrane Library can also be studied. In step four, the file was selected from the *browse* button in the *choose a file* section. Using their plain text format files was recommended to researchers in this step. Plain text format files can be downloaded from WoS or other sources. In the 5th and last step, after clicking the start button, the data was uploaded to the system and became ready for analysis. *Biblioshiny* had a user-friendly interface for bibliometric analysis. All analyses can be done by clicking on the relevant tabs/buttons. An example is given in Figure 6 to show how the analyses were done.

Figure 6

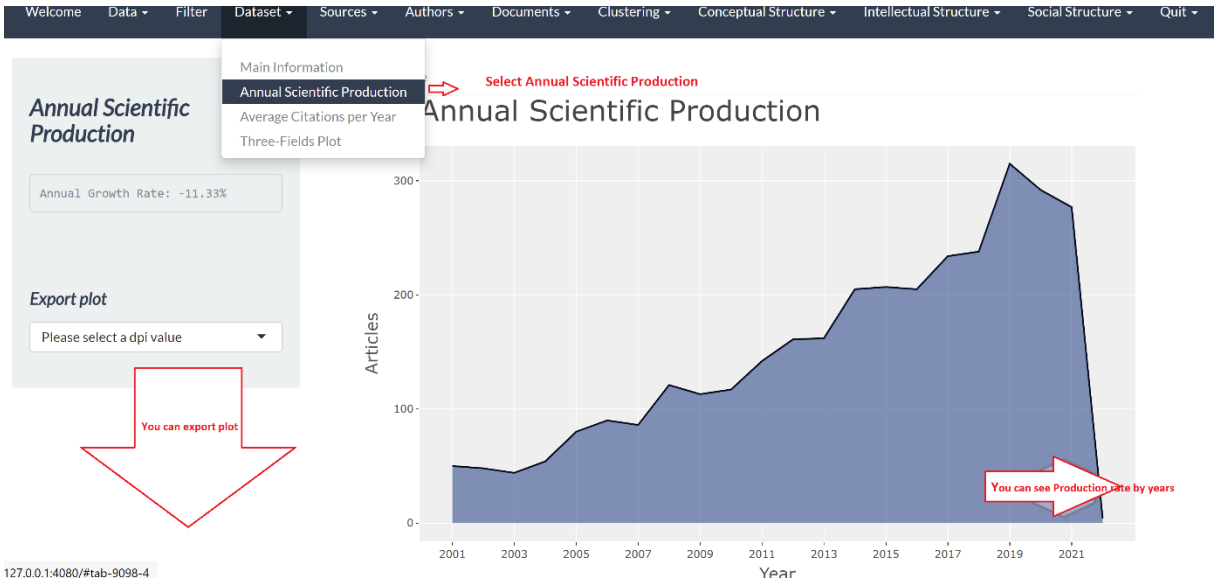
Main Information about the Bibliometric Data



When the general information about bibliometric data was examined, 3388 articles from 639 sources were included in the research. 7672 authors wrote 3388 articles related to IRT. The average citation amount of each document was 21.4. The average citation per year per doc was 2.1. The number of author's keywords was 6433. The annual production number of IRT-related publications was derived from the "annual scientific production" option on the shrinking tab. In addition, *biblioshiny* offered the option to download each plot in different dpis (dots per inch for resolutions). It is possible to export plots from 75 dpi to 600 dpi.

Figure 7

Annual Scientific Production

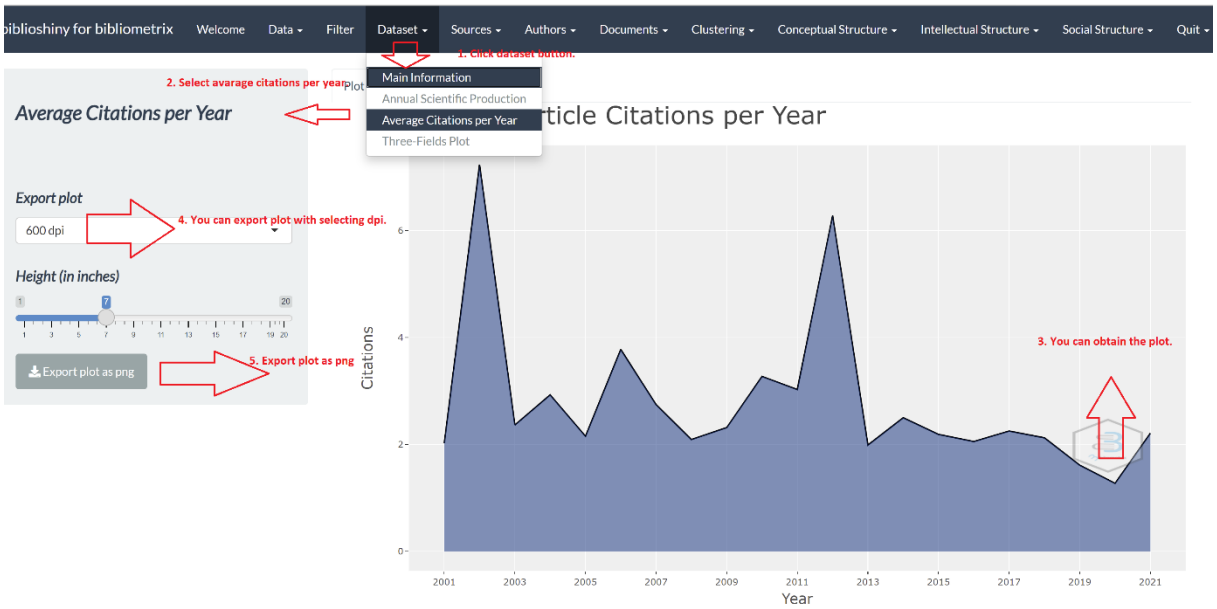


When Figure 7 was examined, the highest number of publications related to IRT was published in 2019 (N= 315). From 2001, the number of publications produced until 2021 was generally increasing. While there were 50 articles searched in WoS related to IRT in 2001, there were 292 articles in 2020. After 20 years, the number of articles increased to 277 in 2021.

A plot in Figure 8 was obtained from the dataset tab by clicking the "*average citations per year*" tab.

Figure 8

Average Article Citations per Year



In Figure 8, the annual citation amounts per article are given. When the plot was examined, the number of annual citations per article in 2002 was at its highest level. Considering the average citations per year, the mean total citations per article was the highest level (N= 144.67) although the number of articles in 2002 was very low (N= 48). Finally, you can see the plot from the Three fields plots option on the dataset tab. At this stage, there were seven steps to follow. After clicking the Three-fields plot tab, the variables and number of items in the middle, left, and right fields must be determined. Once these conditions are set, the apply button must be clicked, and the plot should be obtained. After all these steps, you can download the plot to your computer by clicking the camera-shaped icon on the plot.

Figure 9

Three-Fields Plot

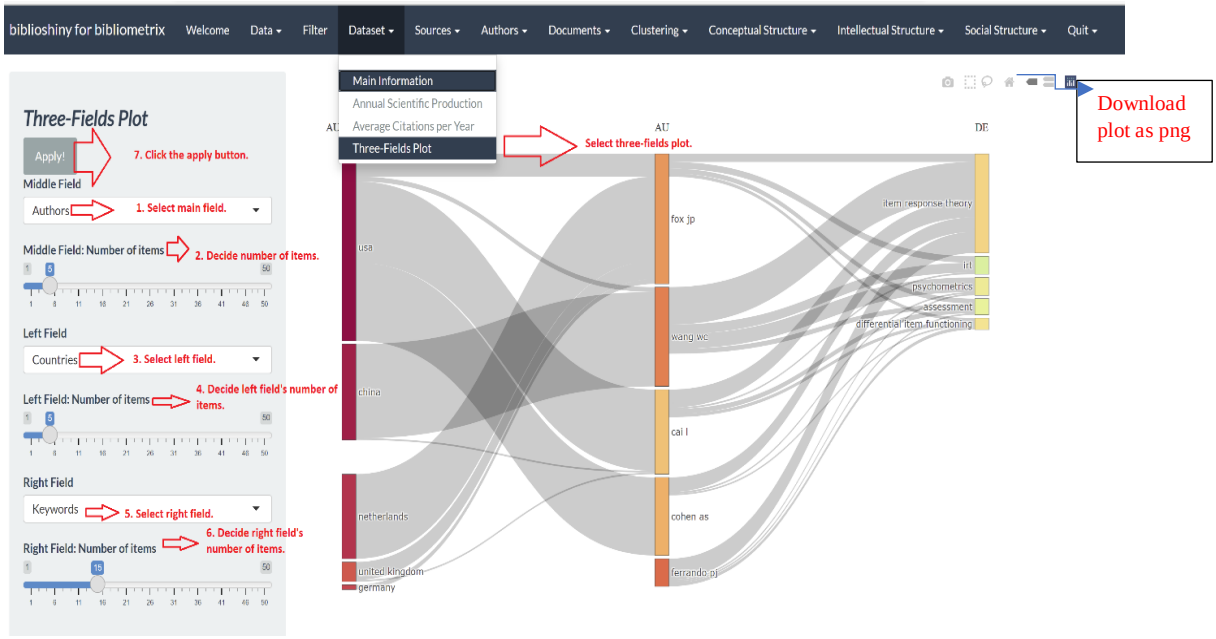


Figure 9 included the Three-Fields Plot. According to the objectives of the researchers in three fields, these fields can be selected from authors, affiliations, countries, keywords, keywords plus, titles, abstracts, sources, references, and cited sources. As part of the research, the author was selected as the middle field, the country as the left field, and the keywords as the right field. The number of items was selected as five for each field. In Table 1, the analysis under all tabs in the *biblioshiny* interface was handled.

Table 1

Sections in Biblioshiny (Adapted from Aria & Cuccurullo, 2017)

Sections	Frames	Bibliometric Technique	Unit of Analysis	Statistical Techniques
Dataset	Main information Annual Scientific production Average Citations per Year Three-Fields Plot			Descriptives

Table 1 (continued)

Sections in Biblioshiny (Adapted from Aria & Cuccurullo, 2017)

Sources	Most relevant Sources Most Local Cited Sources Bradford's Law Source Impact Source Dynamics		Sources	
Documents	Most Global Cited Documents Most Local Cited Documents Most Local Cited References Reference Spectroscopy Most frequent Words WordCloud TreeMap Word Dynamics Trend Topics		Documents, Cited references, Words	
Clustering	Clustering by Coupling			Clustering
Conceptual Structure	Co-occurrence Network Thematic Map Thematic Evaluation Factorial Analysis	Co-word	ID, DE (keywords), TI, AB, Full document	Network Analysis, Factorial Analysis (CA; MICA; MDS), Thematic mapping, Thematic evolution, Topic modeling
Intellectual Structure	Co-citation Network Historiograph	Co-citation, Citation	References, Authors, Journals	Network analysis, Historiograph
Social Structure	Collaboration Network Collaboration WordMap	Collaboration	Authors (co-authorship, Institution, Journal)	Collaboration network

Findings

In this section, the steps and findings of bibliometric analysis with IRT articles were shared.

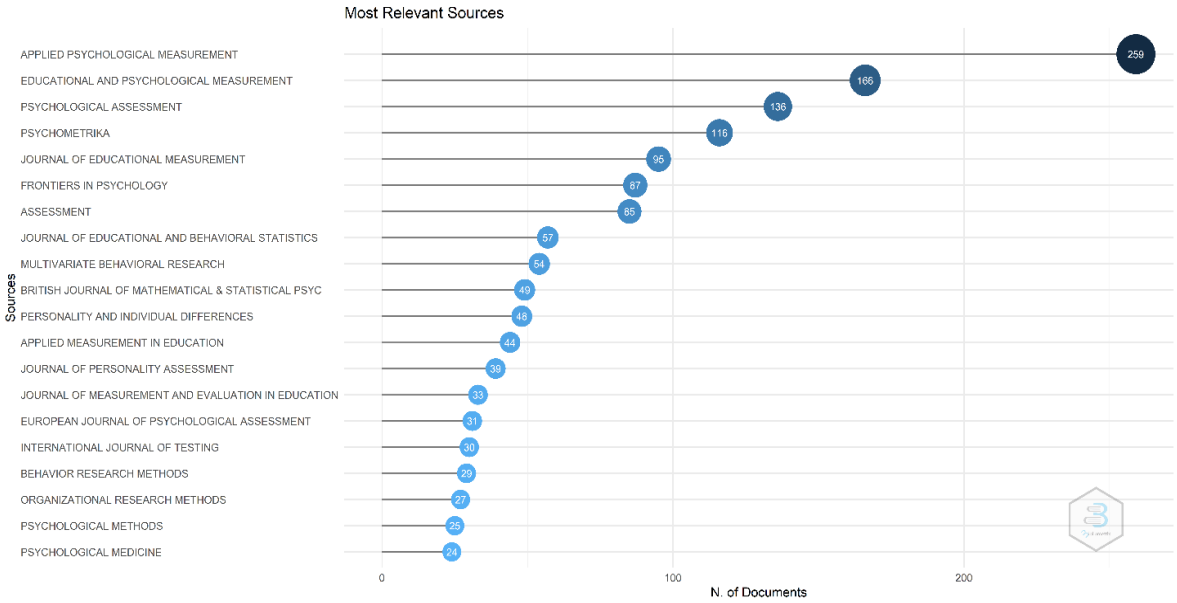
Sources

This section presents the findings of the articles included in this research within the scope of the source (most relevant source, source local impact by h index, source growth, and most local cited sources).

Figure 10 visualized which source most relevantly publishes articles addressing IRT. The Biblioshiny→Source→Most Relevant Source tabs were clicked respectively to obtain the plot in Figure 10.

Figure 10

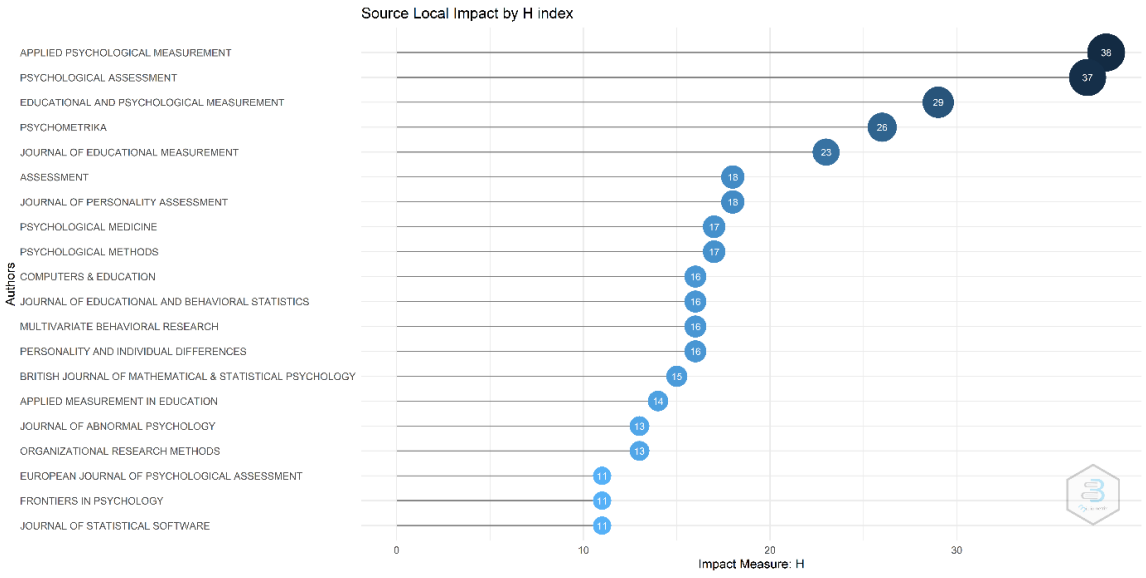
Most Relevant Sources



When the articles on IRT were examined in the field of Education and Psychology, these articles included in this research were most commonly published in Applied Psychological Measurement (N=259), Educational and Psychological Measurement (N=166), Psychological Assessment (N=136), Psychometrika (N=116), Journal of Educational Measurement (n=95). There are 33 publications related to IRT in the Journal of Measurement and Evaluation in Education and Psychology (JMEEP), which ranks 14th. The H index sequences of these journals are shown in Figure 11. The Biblioshiny→Source→Source Impact→H index→Apply tabs were clicked respectively to obtain the plot in Figure 11.

Figure 11

Source Local Impact by H index



Considering the source local impact by H index, Applied Psychological Measurement (H Index= 38), Psychological Assessment (H Index= 37), Educational and Psychological Measurement (H Index=29), Psychometrika (H Index= 26), Journal of Educational Measurement (H index= 23) journals were ranked, respectively (see Figure 11). In Figure 12, the ten journals that published the highest number of articles on IRT are given source growth plot by the year. Biblioshiny→Source→Source Dynamics→Cumulate/Per year→Apply stages have been followed to obtain these plots, respectively.

Figure 12

Source Growth

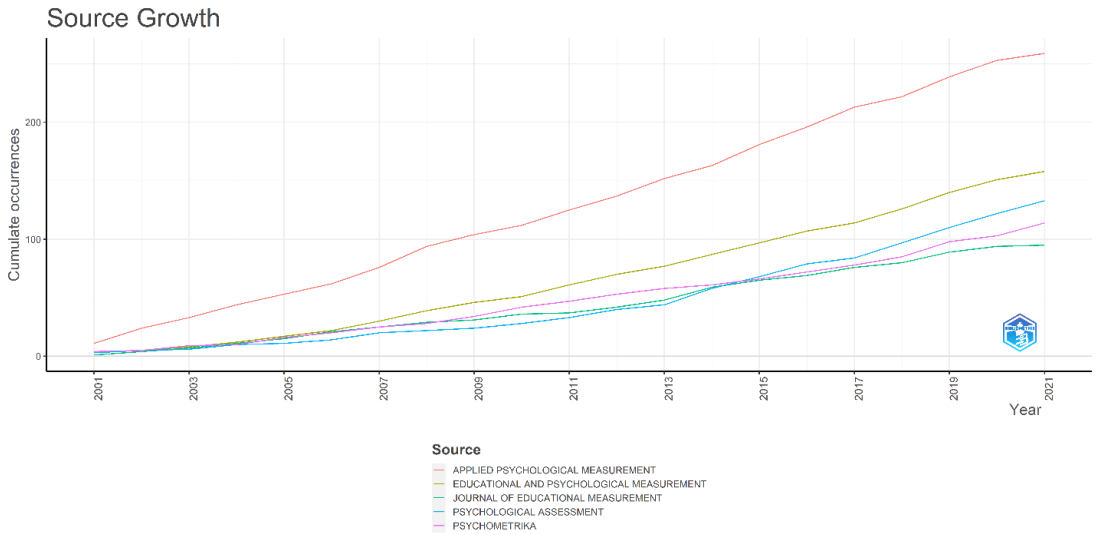
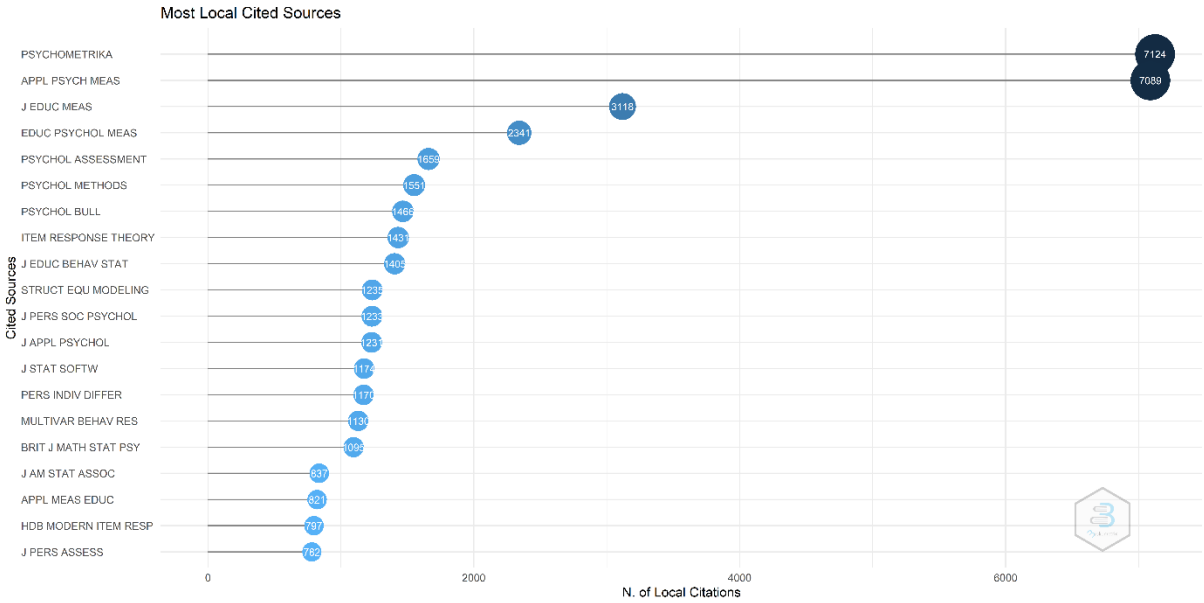


Figure 12 shows the publications by the years given on the left and the developments for each year given on the left. Since 2001, the number of articles related to IRT has been increasing for each journal. For example, while there were 11 articles on IRT in 2001, there were 259 articles in 2021 in the Applied Psychological Measurement Journal. Figure 13 presents the local citation amounts of the 20 most cited journals related to IRT. Biblioshiny→Source→Most Local Cited Sources→Apply stages were followed to obtain the most local cited sources plot, respectively.

Figure 13

Most Local Cited Sources



According to Figure 13, the most local cited sources for IRT-related articles were Psychometrika (Number of local citations/N=7124), Applied Psychological Measurement (N=7089), Journal of Educational Measurement (N=3118), Educational Psychological Measurement (N=2341), and Psychological Assessment (N=1659), respectively.

Authors

The findings of the analyzes of the authors conducting research related to IRT (most relevant authors, top-authors' production over time, most relevant affiliations, corresponding author's country, most cited countries) were handled in this section. Figure 14 provides the number of publications of the twenty authors who have published the highest number of articles about IRT. Biblioshiny→Authors→Most Relevant Authors→N of Documents →Apply stages were followed to obtain this plot respectively.

It was seen that Ferrando PJ had 39 publications in Figure 14. Other most relevant authors are Wang WC (Number of articles=36), Fox JP (N=26), Cohen AS (N= 25), and Cai L (N=24). In Figure 15, these authors' publications and citation amounts were shown by year. To obtain this plot, Biblioshiny→Authors→ Authors' Production over Time→Apply stages were followed respectively.

Figure 14

Most Relevant Authors

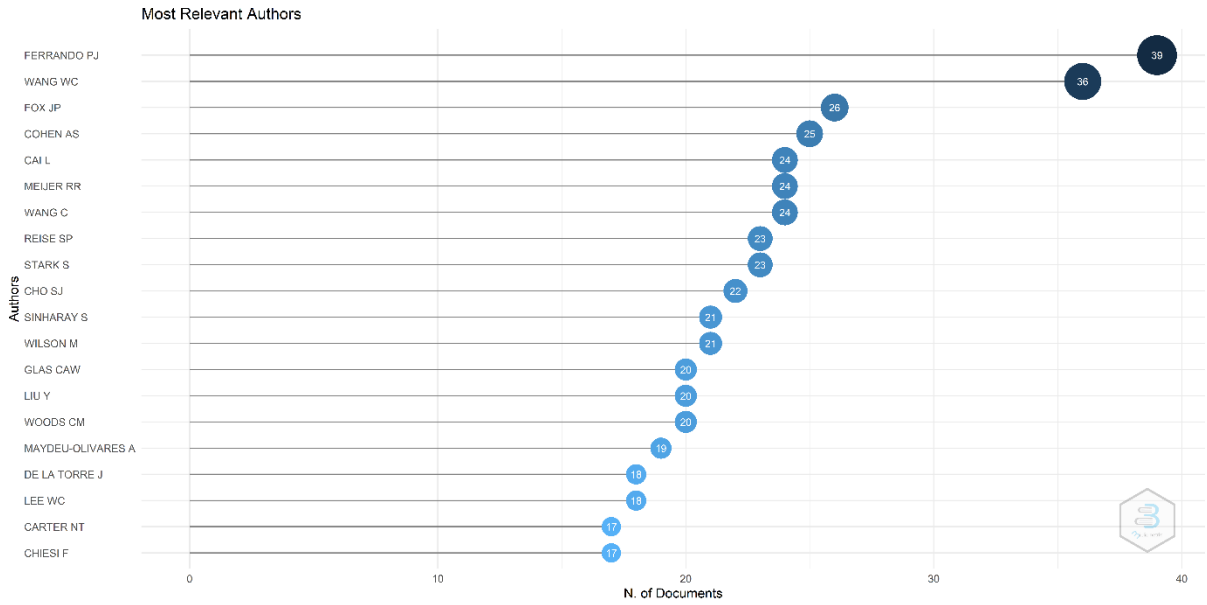


Figure 15

Top-Authors' Production Over the Time

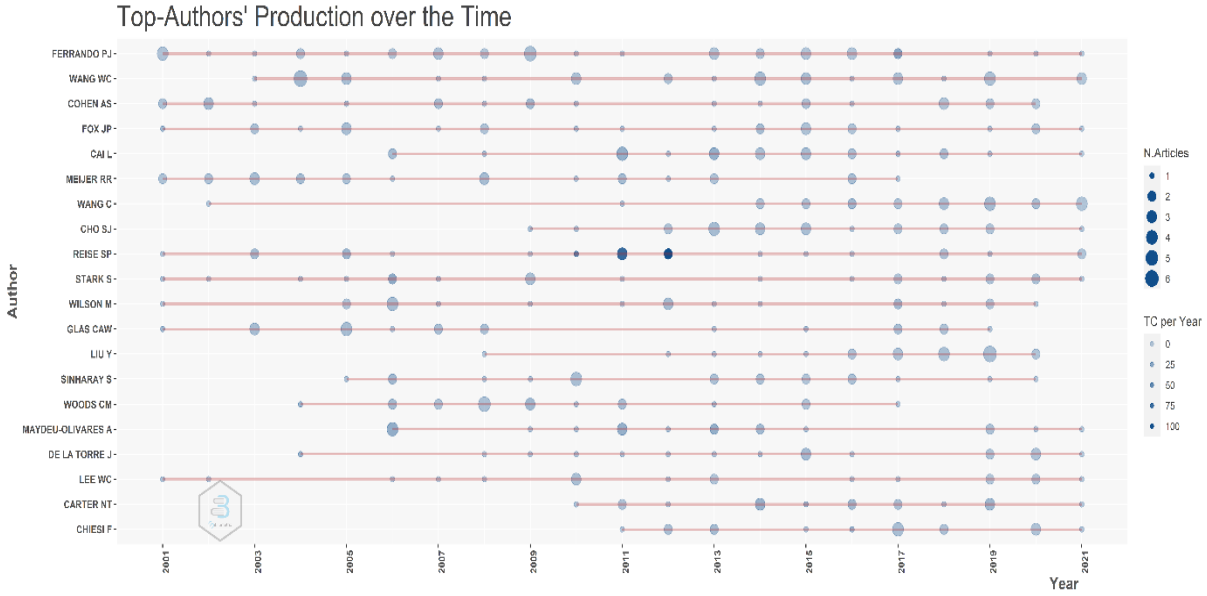
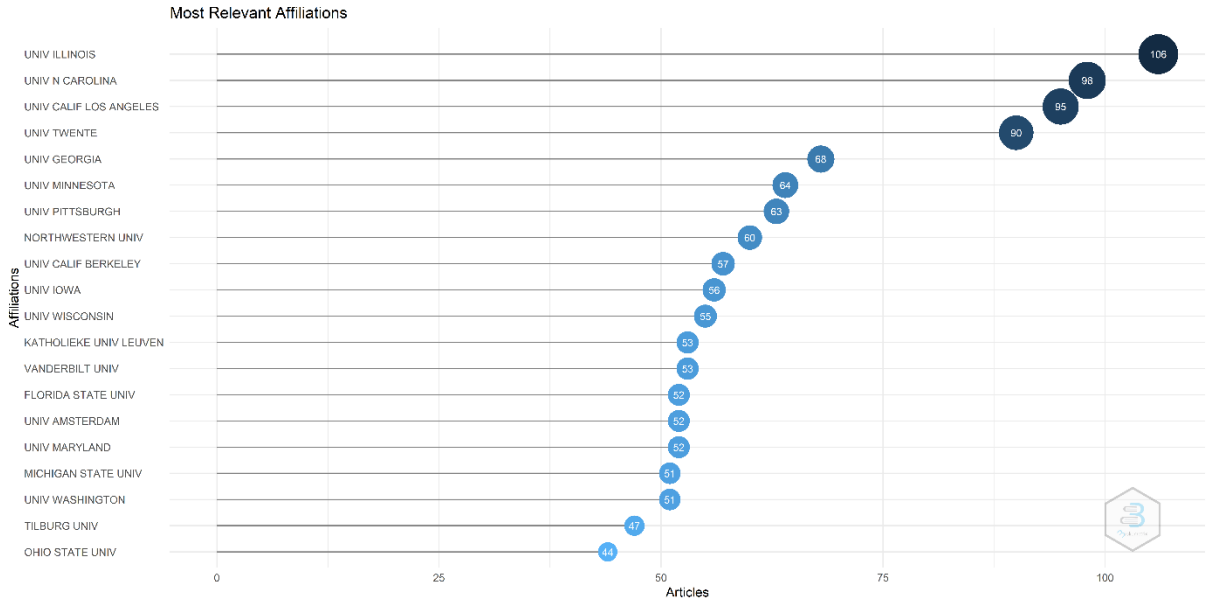


Figure 15 illustrates top-authors' production over time; for example, the circles were darker and larger since Reise SP had more articles and citations in 2011 than in other years. Reise SP had three publications in 2011 and an average of 76.83 citations per year. In Figure 16, the organizations of the authors of the articles were given. To obtain this plot, the Biblioshiny→Authors→ Most Relevant Affiliations→ Affiliation Name Disambiguation-Yes→Apply stages were followed, respectively.

Figure 16

Most Relevant Affiliations



The authors of IRT-related articles were mainly from the University of Illinois Urbana-Champaign (Number of articles= 106), University of North Carolina at Chapel Hill (N= 98), University of California, Los Angeles (N= 95), University of Twente (N= 90), and University of Georgia (N= 68) (see Figure 16). Considering the universities in Turkey within the scope of the research criteria, the most relevant affiliation was Hacettepe University (Number of articles=29) in Turkey. In Figure 17, the authors' countries of the articles were given. Table 2 and the red bar in Figure 17 indicate the multiple country publication (MCP), and the turquoise bar represents the single country publication (SCP). To obtain this plot, Biblioshiny→Authors→ Corresponding Author's Country→Apply stages were followed respectively.

Figure 17

Corresponding Authors' Country

Figure 17 states that the corresponding authors were most commonly found in the USA. China (Number of articles= 200), Netherlands (N= 184), Germany (N= 161), and Spain (N=138) ranked the top five corresponding author's countries. Turkey, on the other hand, was the 9th country with a corresponding authors rate who conducted IRT studies. Table 2 provides the number, ratio, and frequencies of IRT-related publications according to the country and the status of being single country publications (SCP) or multiple-country publications (MCP).

Table 2

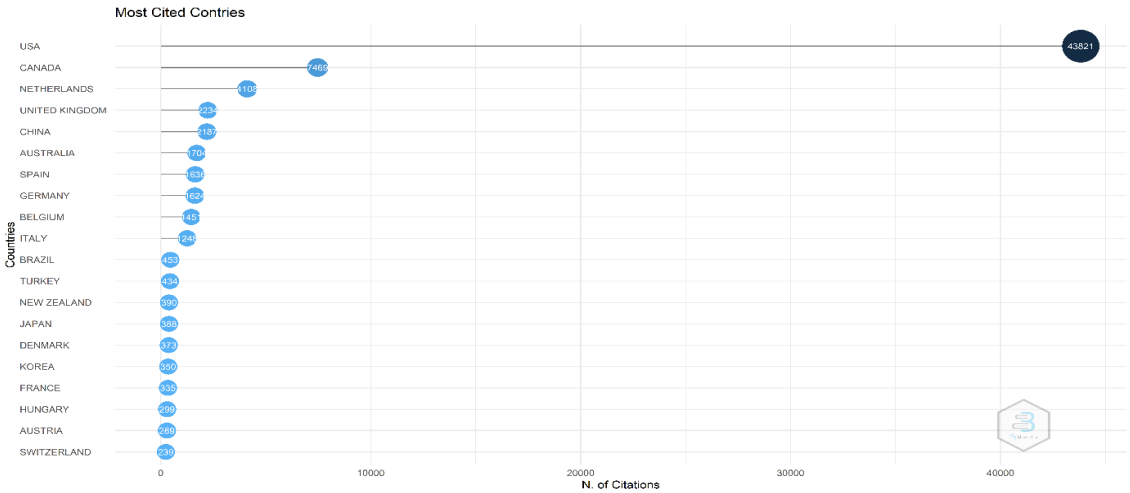
Corresponding Author's Country

Country	Articles	Frequency (f)	SCP	MCP	MCP_Ratio
USA	1605	0.47	1378	227	0.14
China	200	0.06	151	49	0.25
Netherlands	184	0.05	145	39	0.21
Germany	161	0.05	120	41	0.25
Spain	138	0.04	114	24	0.17
Italy	105	0.03	77	28	0.27
Canada	102	0.03	58	44	0.43
United Kingdom	94	0.03	46	48	0.51
Turkey	93	0.03	84	9	0.10
Australia	90	0.03	61	29	0.32

When the corresponding author's country was analyzed, the USA ranked first with 1605 articles (f= 0.47). Table 2 shows that authors mostly wrote the articles in the USA from a single country (N= 1378), but 227 studies were conducted in collaboration with authors from other countries. The number of articles on IRT in the USA was about eight times higher than the number of articles in China, which ranks second. As can be seen from Table 2, for example, there were 93 articles (searched in SSCI, SCI-E, ESCI, and A&HCI) in Turkey according to the criteria set out in WoS. Only 9 of these articles were multiple-country studies. The most cited countries are shown in Figure 18. To obtain this plot, the Biblioshiny→Authors→ Most Cited Countries→Measure-Total citations→Apply stages were followed respectively.

Figure 18

Most Cited Countries



In Figure 18, the most cited countries were the USA (Number of citations= 43821), Canada (N= 7469), Netherlands (N=4108), United Kingdom (N= 2234), and China (N= 2187), respectively. A total of 43821 citations were made to articles in the USA, compared to 27.303 citations per article. Although the total number of citations in Canada, which is in second place, was approximately one-sixth of that of the USA, the average number of article citations in Canada had the highest value with 73.23. When these numbers were examined for Turkey, there were 434 total citations, while the average number of article citations was 4.667.

Social Structure

The collaboration network was shown in this section.

Figure 19

Collaboration Network

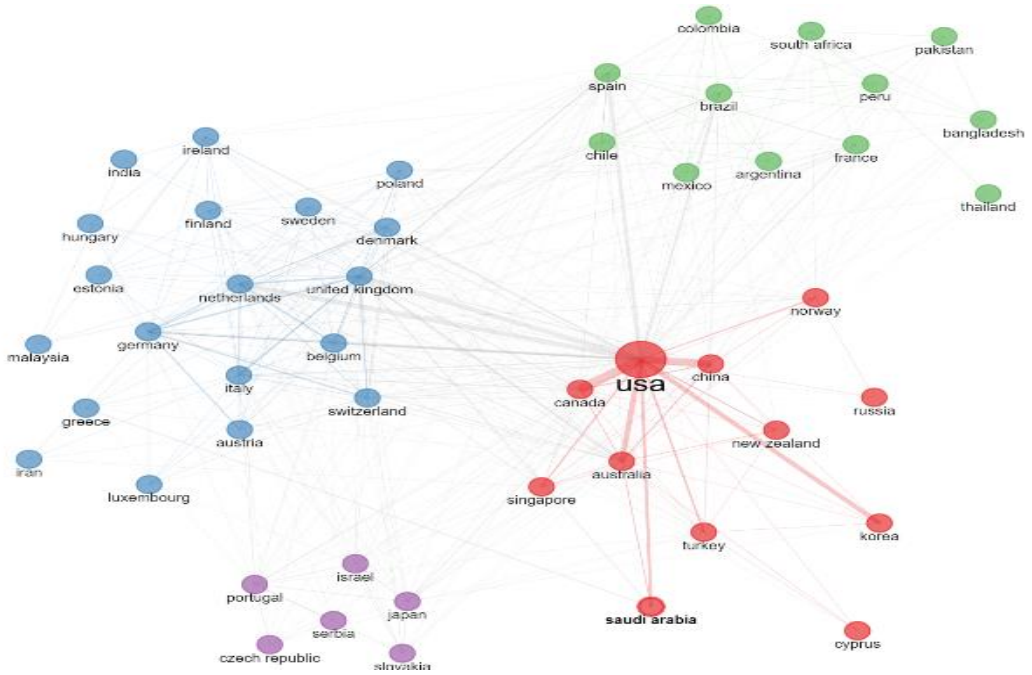


Figure 19 displays a collaboration network for countries. To obtain this figure, the Biblioshiny→Social Structure→ Collaboration Network→Field-Countries→Apply stages were followed respectively. In this analysis, we used 50 number of nodes, we chose Louvain clustering algorithm, and used association as normalization.

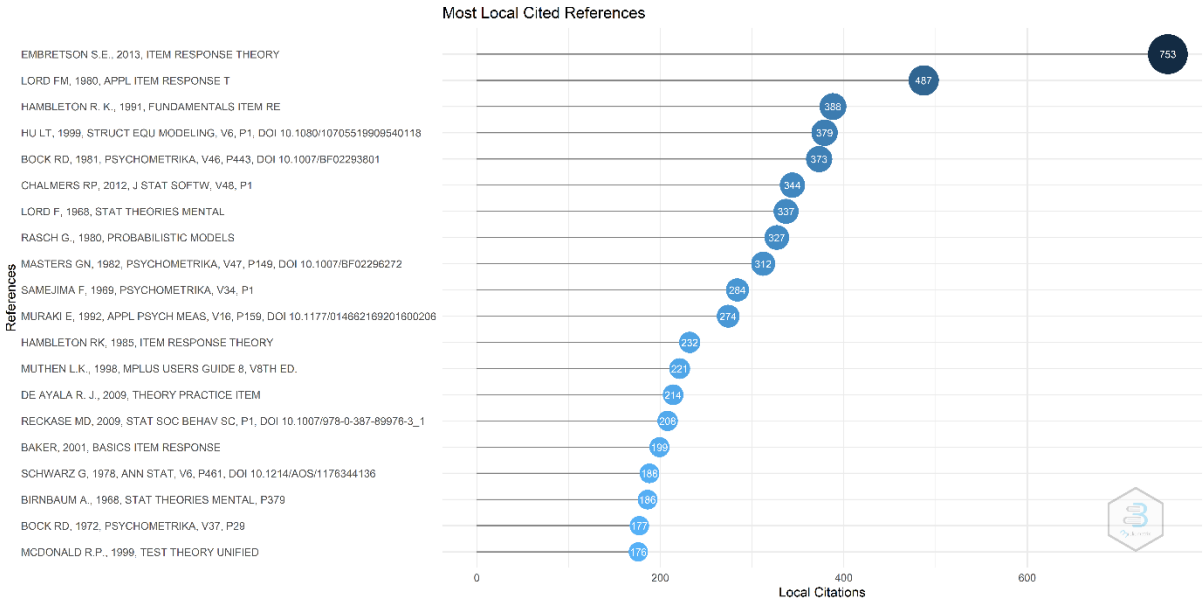
As can be seen in Figure 19, there were five clusters. Each cluster is represented by a color. It can be interpreted that the countries in the same cluster were in cooperation. These clusters were shown in different colors. USA, China, Canada, Australia, Turkey, Korea, Norway, New Zealand, Singapore, Russia, Saudi Arabia, and Cyprus were among the first cluster. Netherlands, Germany, United Kingdom, Italy, Belgium, Switzerland, Austria, Denmark, Sweden, Finland, Poland, Iran, Greece, Ireland, India, Estonia, Hungary, Luxembourg, and Malaysia were among the second cluster. Spain, Brazil, France, Argentina, Peru, Chile, Colombia, South Africa, Mexico, Thailand, Pakistan, and Bangladesh were among the third cluster. Japan, Portugal, Czech Republic, Serbia, Slovakia, and Israel were among the fourth cluster.

Documents

Most local cited references, reference spectroscopy, most frequent words, tree map, word dynamics, and trend topics (based on keyword-plus and author's keyword) were shown in this section. In Figure 20, according to the inclusion criteria in WoS, most local cited references were listed. To obtain Figure 20, the Biblioshiny→Documents→ Most Local Cited References→Apply stages were followed respectively.

Figure 20

Most Local Cited References



Regarding Figure 20, the greatest number of citations were done to Embretson's and Reise's Item Response Theory (2013) (N= 753) in the research included in this bibliometric research. Then the most local references were publications of Lord FM, 1980 (N= 487), Hambleton R. K., 1991 (N=388), Hu LT, 1999, (N= 379), Bock RD, 1981, (N= 373), Chalmers RP, 2012 (N= 344), Lord F, 1968 (N= 337), Rasch G., 1980 (N= 327), Masters GN, 1982 (N=312), and Samejima F, 1969, (N= 284). Reference publication year spectroscopy is given in Figure 21. To obtain this plot, the Biblioshiny→Documents→ Reference Spectroscopy→Apply stages) were followed respectively.

Figure 21

Reference Publication Year Spectroscopy

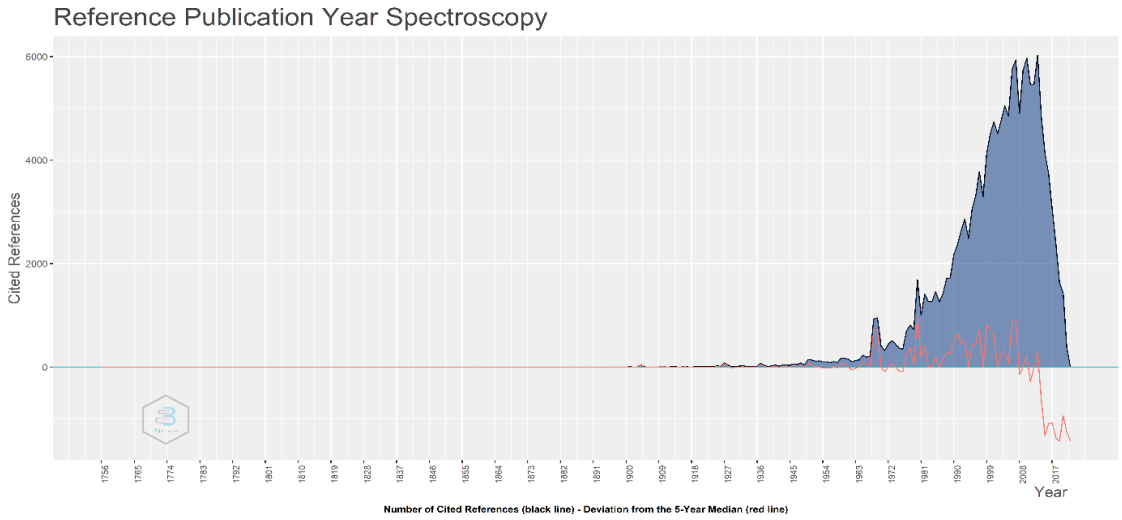
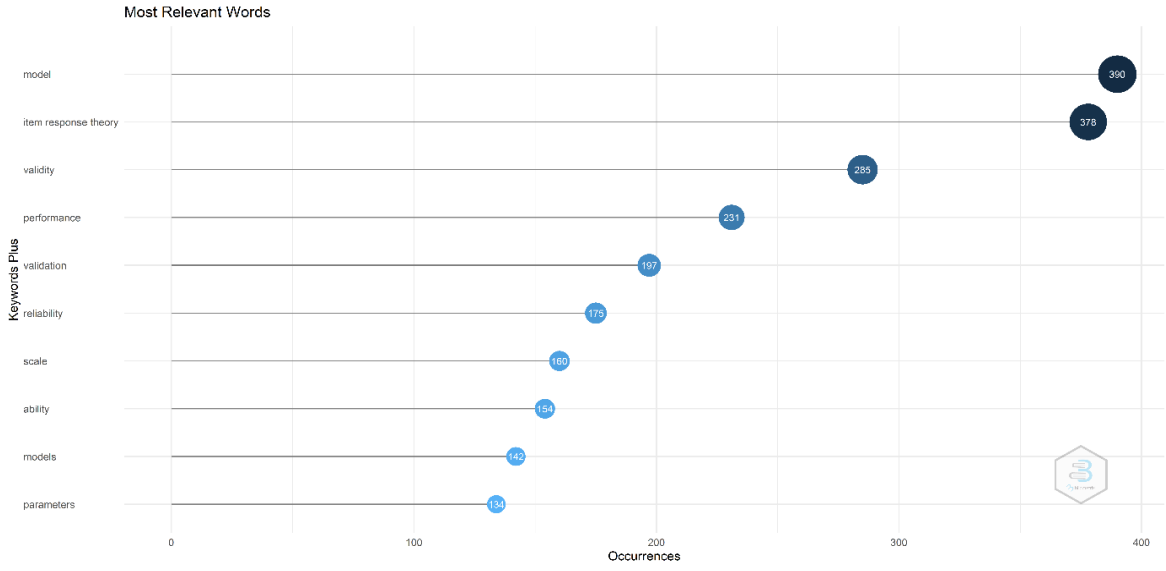


Figure 22 gives the most relevant words. To obtain this plot, the Biblioshiny→Documents→ Most Frequent Words→Field-Keywords plus→Apply stages were followed respectively.

Figure 22

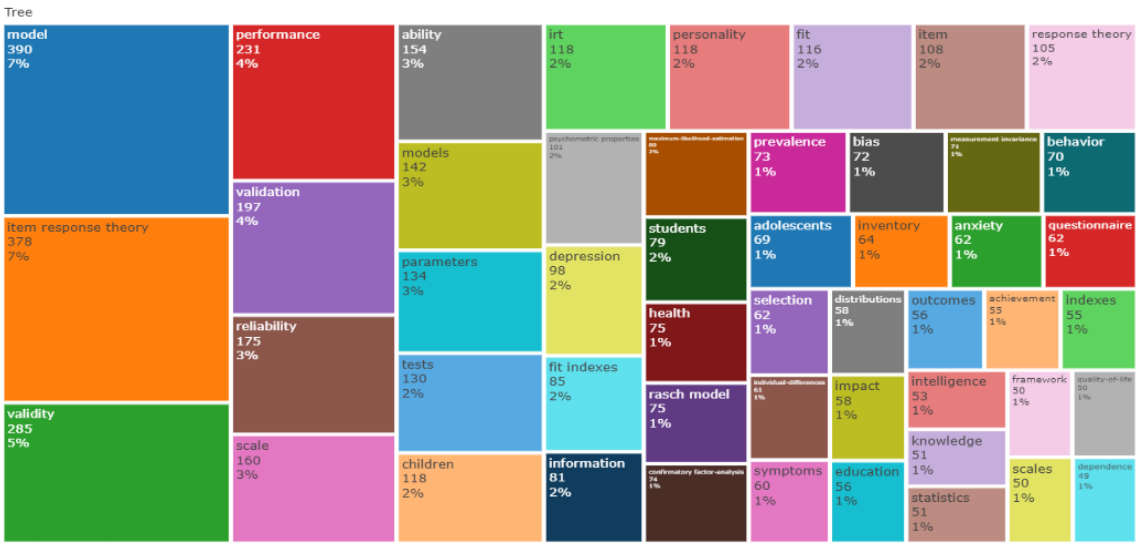
Most Relevant Words



When the articles including in this research were examined, it was seen that keywords such as model (f= 390), item response theory (f=378), validity (f=285), performance (f=231), validation (f= 197), reliability (f=175), scale (f= 160), ability (f= 154), and parameters (f= 134) were used (see Figure 22). Figure 23 gives tree map based on keywords. To obtain this plot, the Biblioshiny→Documents→ TreeMap→Field-Keywords plus→Apply stages were followed respectively.

Figure 23

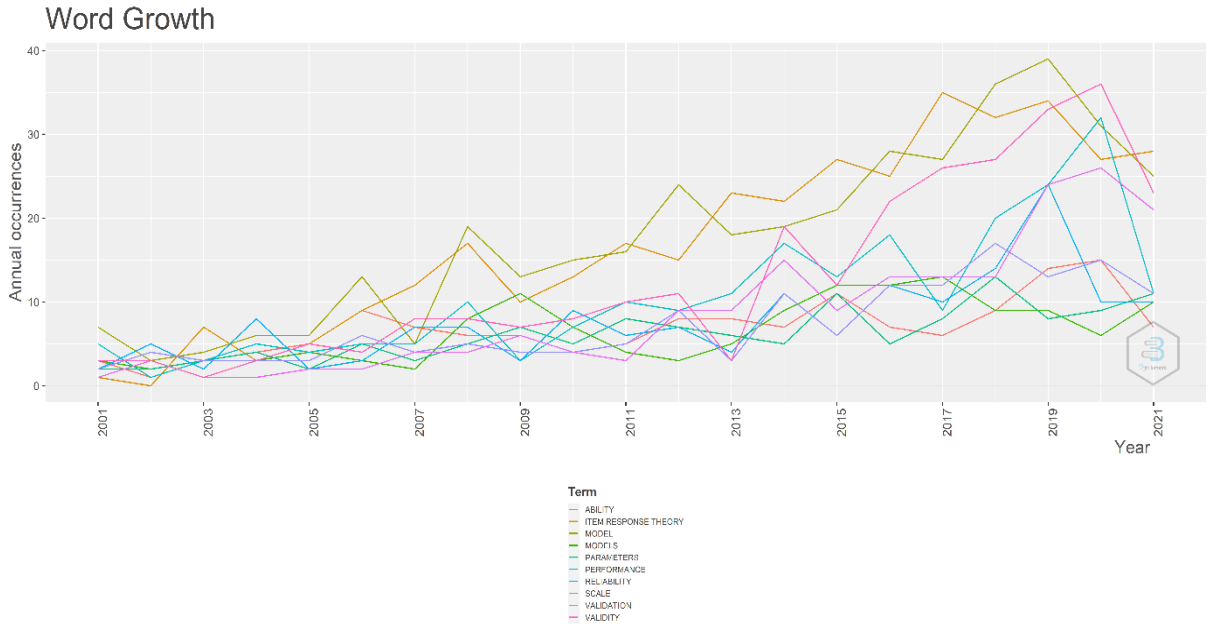
Tree Map Obtained from Keywords Plus



When Figure 23 was examined, it was seen that concepts such as validity (N=285, %5), performance (N=231, %4), validation (N=197, %4), and reliability (N=175, %3) stood out in the articles included in the analysis. The frequency with which these concepts have been used in articles over the years is given in Figure 24 with the word growth plot. This plot was obtained by following the Biblioshiny→Documents→ Word Dynamics→Field-Keywords plus→ Occurrences-Per Year→Apply stages respectively.

Figure 24

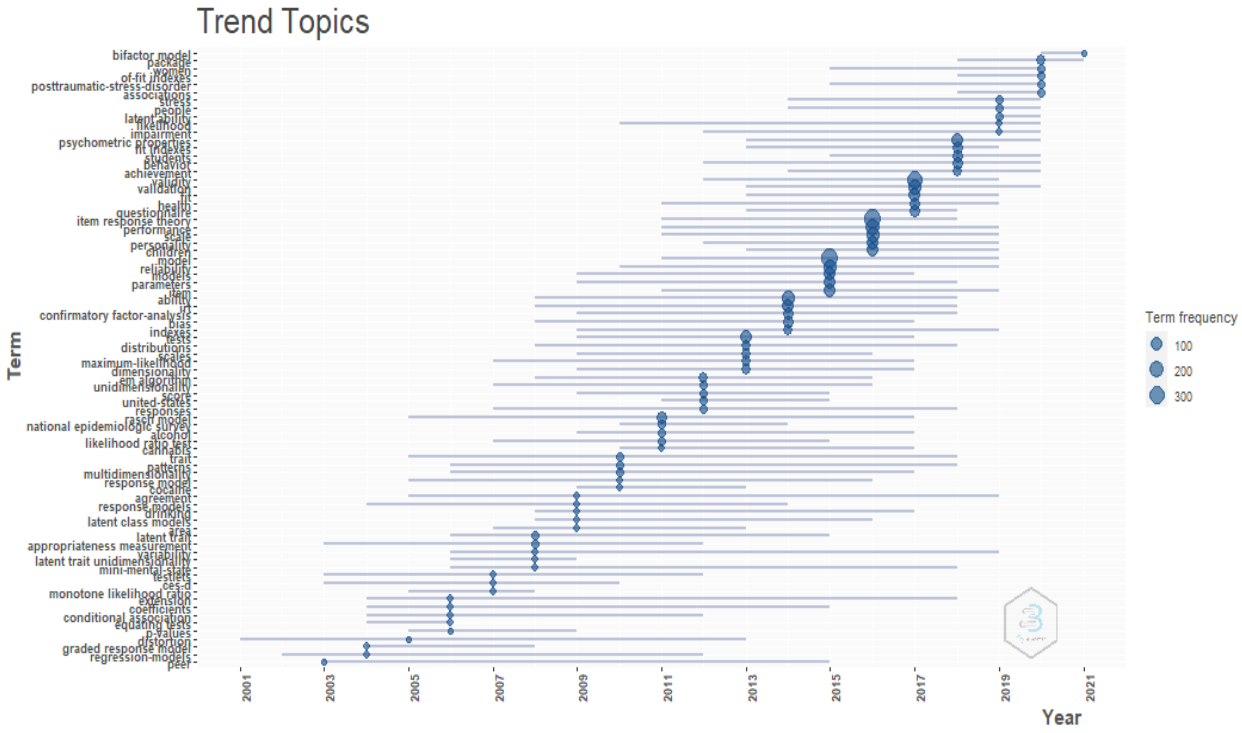
Word Growth



When Figure 24 was considered, for example, it was seen that the concept of validity was included three times in 2001, 27 times in 2018, 33 times in 2019, 36 times in 2020, and 23 times in 2021. Figure 25 is given a trend topics plot. "Word minimum frequency" parameter was taken as five and the "number of words per year" parameter was taken as three when constructing trend topics plots. This plot was obtained by following the Biblioshiny→Documents→ Trend Topics→Field-Keywords Plus →Apply stages respectively.

Figure 25

Trend Topics Based on Keywords Plus

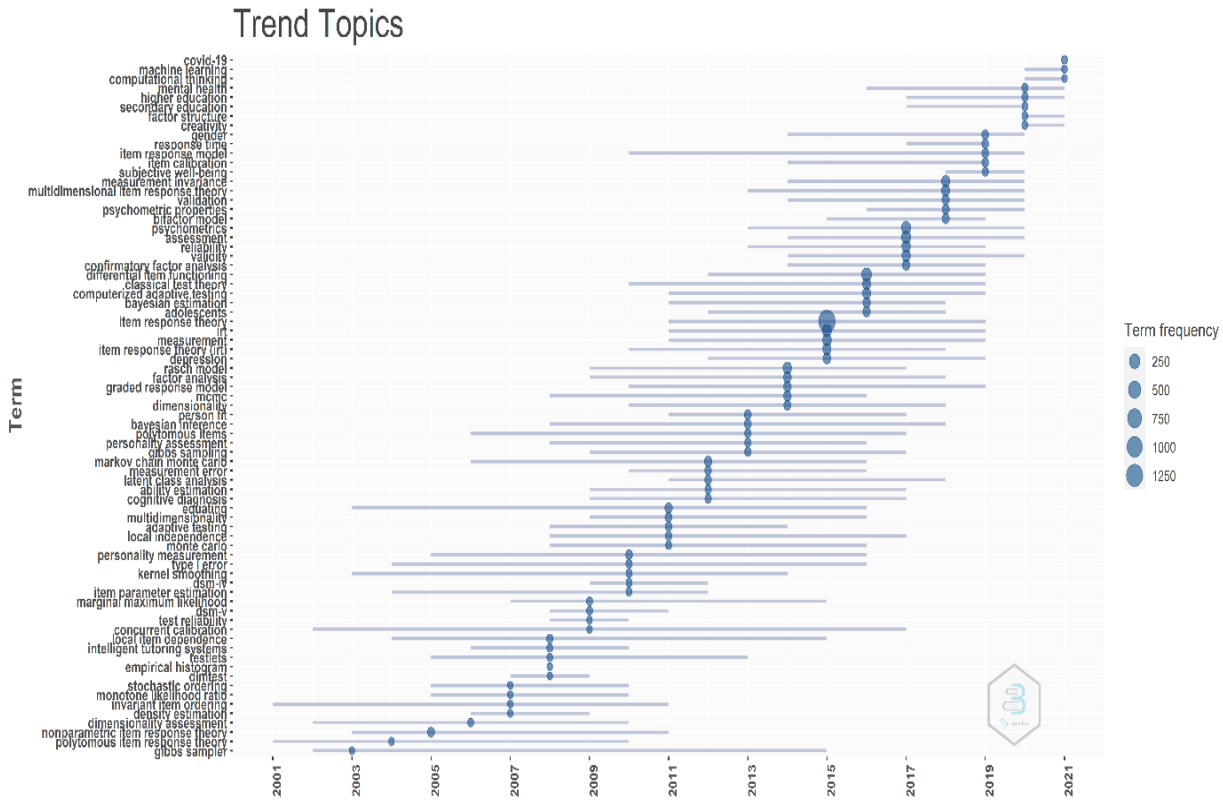


When Figure 25 was examined, it had been seen that the most bi-factor models have been trending topics based on keywords plus in recent years. In other words, the bi-factor model has been studied in research on IRT in recent years. In the trend topics plot, the size of the circles shows the frequency of the term and the length of the lines shows how long it has been studied. In Figure 26, the Trend topics plot was given based on authors' keywords. This plot was obtained by following the Biblioshiny→Documents→ Trend Topics→Field-Author's Keywords →Apply stages respectively.

It is seen that COVID-19 has been a trending topic based on author's keywords in recent years (see Figure 26). COVID-19 was a trending topic even in research on IRT. In recent years, two other trend topics in IRT research are machine learning and computational thinking based on author's keywords. The most commonly used author's keywords were ranked as item response theory (f=1279), differential item functioning (f= 185), psychometrics (f= 137), assessment (f= 135), IRT (f= 118), measurement (f= 105), reliability (f= 94), Rasch model (f= 89), measurement invariance (f= 74), multidimensional item response theory (f= 72), classical test theory (f= 68), computerized adaptive testing (f= 65)), factor analysis (f= 50), graded response model (f= 41), Markov Chain Monte Carlo (f= 39), equating (f= 36), and validation (f= 34).

Figure 26

Trend Topics Based on Author's Keywords

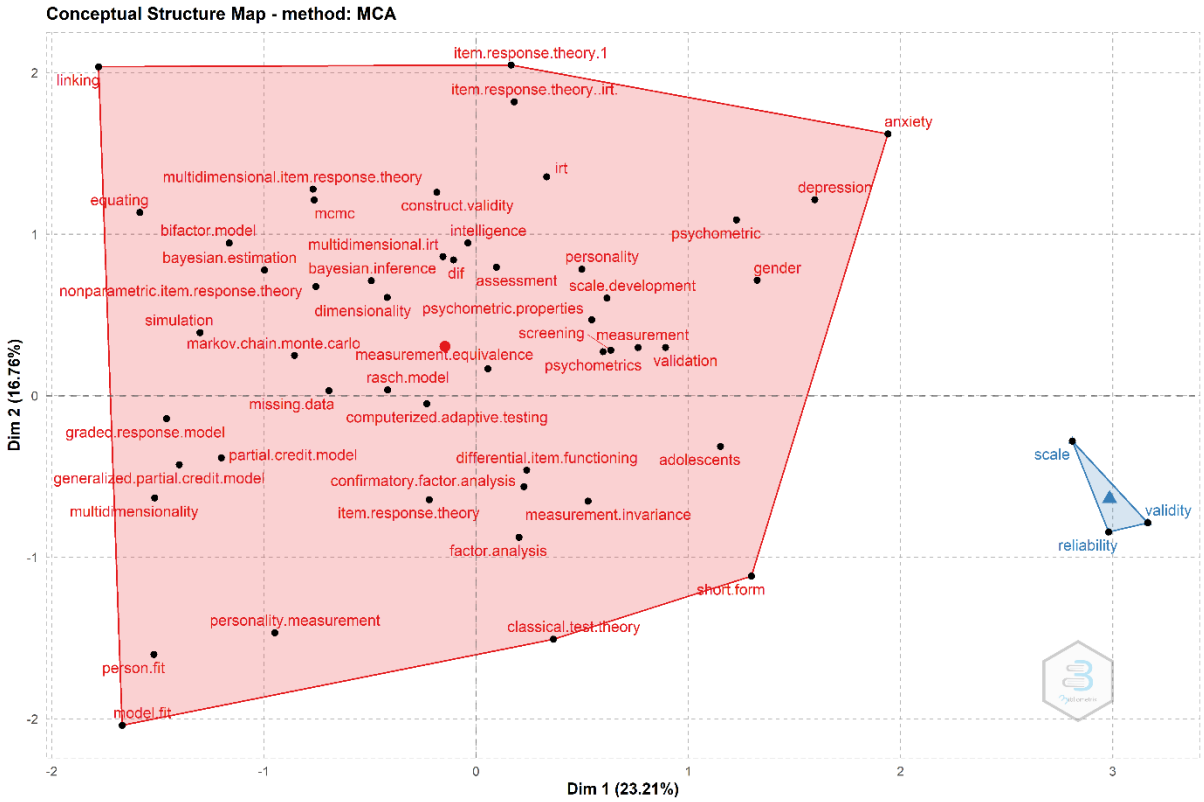


Conceptual Structure

This section displays a conceptual structure map, topic dendrogram, network map based on author's keywords, and thematic map. Figure 27 was illustrated a conceptual structure map. This plot was obtained by following the Biblioshiny→Conceptual Structure→ Factor Analysis→Method-Multiple Correspondence Analysis→Field-Author's Keywords →Apply→Word Map stages respectively.

Figure 27

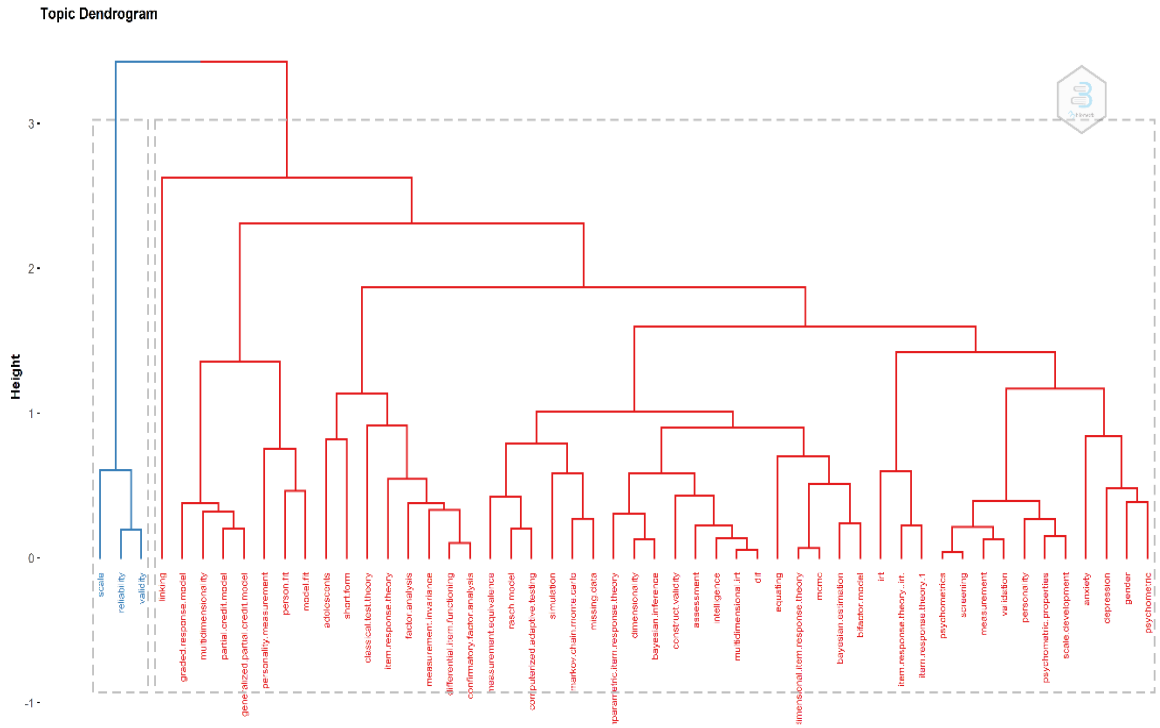
Conceptual Structure Map



As a result of the factorial analysis, when the keywords of the articles about IRT are examined in this study, the following concepts are placed in the first cluster with a high factor load in the first dimension: item response theory, differential item functioning, psychometrics, assessment, IRT, measurement, Rasch model, measurement invariance, multidimensional item response, classical test theory, computerized adaptive testing, depression, factor analysis, graded response model, confirmatory factor analysis, bayesian estimation, Markov Chain Monte Carlo, equating, MCMC, adolescents, validation, personality, partial credit model, psychometric properties, bi-factor model, missing data, scale development, simulation, dimensionality, nonparametric item response theory, linking, short form, measurement equivalence, construct validity, multidimensionality, personality measurement, screening, anxiety, gender, generalized partial credit model, person fit, bayesian inference, intelligence, model fit, psychometric keywords were collected in a single factor. Scale, reliability, and validity were included in a second dimension/cluster (see Figure 27). The dendrogram is given in Figure 28. This figure was obtained by following the Biblioshiny→Conceptual Structure→ Factor Analysis→Method-Multiple Correspondence Analysis→Field-Author's Keywords →Apply→Topic Dendrogram stages, respectively.

Figure 28

Topic Dendrogram



When the Topic Dendrogram was examined, the distance between the clusters was seen on the y-axis of the dendrogram. On the X-axis, there were subject concepts of the data points that make up the clusters. The keyword network map is given in Figure 29. This figure was obtained by following the Biblioshiny→Conceptual Structure→ Co-occurrence Network→Field-Author's Keywords →Apply→Network Map stages respectively.

Regarding the network map based on author's keywords; item response theory, differential item functioning, psychometrics, assessment, measurement, Rasch model, measurement invariance, item response theory(IRT), classical test theory, computerized adaptive testing, graded response model, confirmatory factor analysis, validation, personality, partial credit model, psychometric properties, data missing, scale development, simulation, dimensionality, nonparametric response item theory, construct validity, multidimensionality, screening, gender, generalized partial credit model, person fit, model fit keywords were found in the first cluster (see Figure 29). Different colors were used for each cluster. When Figure 29 was considered, there were 50 concepts and 6 clusters. Keyword thematic map based on author's keyword was given in Figure 30. This figure was obtained by following the Biblioshiny→Conceptual Structure→ Thematic Map→Field-Author's Keywords →Apply→Map stages, respectively. The number of words was taken as 6433, minimum cluster frequency (per thousand docs) was taken as 20. Louvain clustering algorithm was selected for the thematic map.

Figure 29

Network Map Based on Author's Keywords

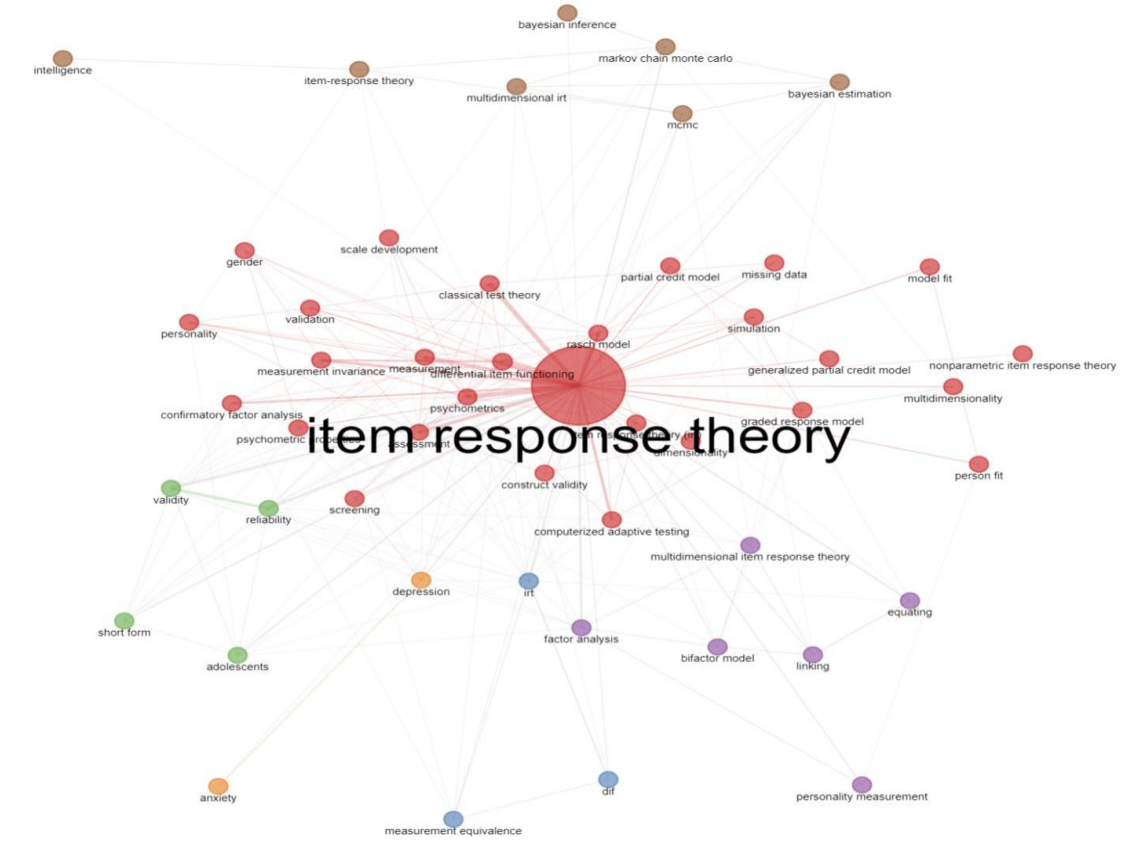
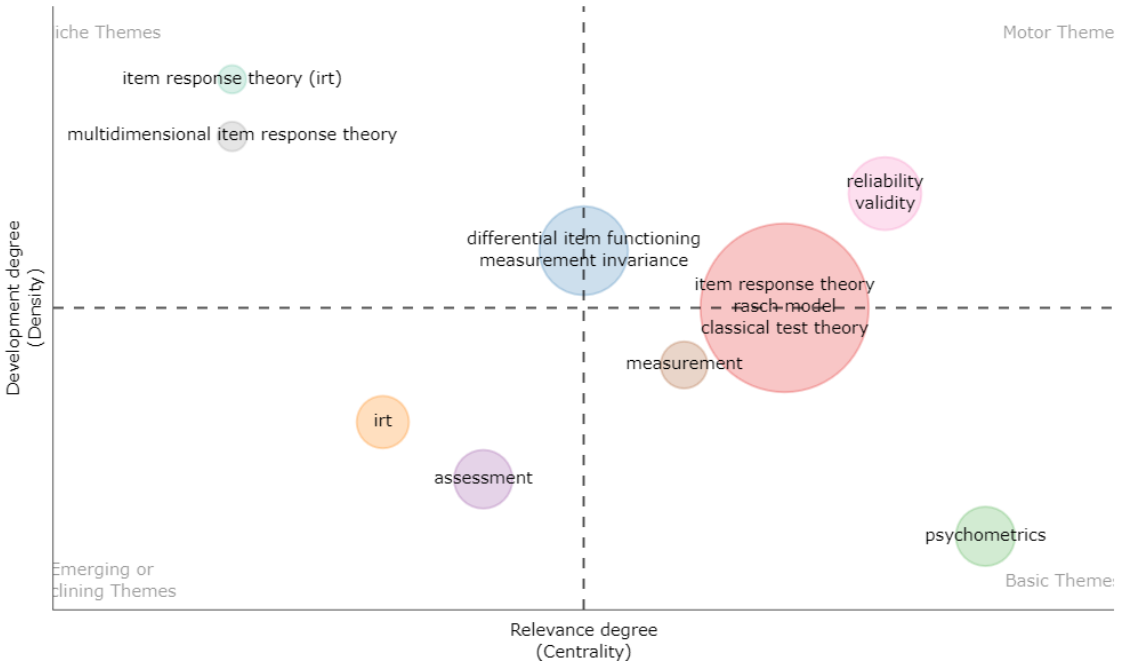


Figure 30

Thematic Map



The thematic map had eight clusters. When the author's keywords were considered, niche themes were item response theory (IRT) and multidimensional item response theory. Emerging themes were IRT and assessment. The Basic (developing) theme was measurement and psychometrics. Motor (developed) themes were reliability and validity. Differential item functioning and measurement invariance cluster was located between two quadrants such as item response theory, Rasch model, and classical test theory cluster.

Conclusion and Discussion

This research aimed to introduce the *biblioshiny* interface opened with the *bibliometrix* package in R programming language and to perform a bibliometric analysis. Publications related to IRT in education and psychology were handled in this research. Between 2001 and 2021, 3388 articles searched in WoS were taken into consideration.

It had been observed that the IRT-related articles included in this research in the field of education and psychology were mostly published in Applied Psychological Measurement, Educational and Psychological Measurement, Psychological Assessment, Psychometrika, and Journal of Educational Measurement, respectively. There were 33 articles on IRT included in this research in the 14th-ranked Journal of Measurement and Evaluation in Education and Psychology (JMEEP). The most cited journals for IRT-related articles were Psychometrika, Applied Psychological Measurement, Journal of Educational Measurement, Educational Psychological Measurement, and Psychological Assessment, respectively. Aksu and Güzeller (2019) conducted a bibliometric analysis of 1367 IRT-related studies searched in WoS between 1980 and 2018 with Citespace II. They found similar results in this research.

According to the determined criteria, when the authors who have published the most articles on IRT was examined, Ferrando PJ had the most publications. When the institutions of the authors who published the most articles on IRT were examined, the most relevant articles were written by authors from the University of Illinois Urbana-Champaign in the USA. It has been observed that employees in institutions in Turkey publish fewer articles. However, the number of corresponding author's articles in Turkey was in the top 10. Increasing the quantity and improving the quality of IRT-related publications is only possible with good cooperation. It was seen that only nine of the articles included in this research were in cooperation with other countries in Turkey. Collaboration with authors from other countries should be increased to improve studies in Turkey. When the total number of citations was examined, it was noteworthy that the number of citations of the articles written in institutions in the USA was approximately 100 times more than the number of articles written from institutions in Turkey. Gómez Benito et al. (2005), in their bibliometric research on 271 articles containing the terms "differential item functioning," "DIF," or "item bias" between 1975 and 2000, published in USA, Holland, Spain, and Canada mostly. They found that authors from these countries were the most productive. Consistent with this research in the article of Gómez Benito et al. (2005), the USA was the most productive corresponding author's country at a rate of 64.5%. Aksu and Güzeller (2019) found that the corresponding author's countries making the biggest contribution to IRT literature were USA, Netherland, Canada, Spain, and China, respectively. When corresponding authors' country and most cited countries were examined, most studies were conducted by authors in the USA. The reason for this situation can be explained by the fact that the USA is a pioneer in the world in the number and quality of universities. According to the Fall 2020 National Center for Education Statistics data, there were approximately 3773 degree-granting postsecondary institutions in the US (see National Center for Education Statistics (NCES), 2020). When these colleges and universities were taken into consideration, it was seen that most of them were research universities where a high number of research was conducted (see NCES, 2020). Authors in Turkey can conduct their research by collaborating with authors from the USA, Canada, and the Netherlands, which are the most cited countries.

When the collaboration network was examined, there were five clusters. USA, China, Canada, Australia, Turkey, Korea, Norway, New Zealand, Singapore, Russia, Saudi Arabia, and Cyprus belong to the first cluster. When the centrality levels were examined, it was seen that the studies included in the research

were mostly from the USA. The reason for the most cited and relevant articles on IRT in the USA may be due to the high number of qualified higher educational institutions (see NCES, 2020). Turkey needs to increase the number of publications to reach a central position. The policies and incentives of institutions such as the Higher Education Institution and The Scientific and Technological Research Council of Turkey (TÜBİTAK) are important in increasing the quality and number of publications. The motivation of the authors in Turkey to make quality publications in journals with high impact factors should be increased with incentives. In bibliometric studies conducted in different countries, it has been stated that researchers working in western countries have difficulties in producing quality publications due to project support, course load, and language deficiencies (Gümüş et al., 2019; Hallinger & Hammad, 2019). Hallinger and Hammad (2019) similarly stated that collaborations and scholarship opportunities are important factors for researchers to produce. Gülmez et al. (2020), in their bibliometric research on educational research, revealed that the visibility of Turkish-based research in foreign-sourced journals was low. They mentioned the limitations in the time, language, and funding resources allocated to the research as the reason for this. In order to overcome these limitations, researchers should be supported in foreign language education, translation, and academic writing. They stated that policymakers and university representatives should support researchers to make international collaborations. On the other hand, Rey-Rocha et al. (2002) indicated that it should not be overlooked that an excess of quantity does not mean an increase in the quality of studies.

In the research included in this bibliometric research, the highest number of references were given to Embretson, S. E., & Reise, S. P.'s (2013) Item Response Theory book. Aksu and Güzeller (2019) found that the most cited authors in the field of IRT were De Ayala, Embretson, Reckase, Reise, and Chalmers, respectively. This finding is partially consistent with this research. When the articles searched in WoS in the field of education and psychology related to IRT were examined, it was seen that keywords such as model, validity, performance, validation, reliability, scale, ability, and parameters were used. Concepts such as validity, performance, validation, and reliability were seen to stand out in the articles included in the analysis. When trending topics based on keywords plus were examined, it was seen that the most bi-factor models have been trending topics in recent years. Trending topics based on the author's keywords plot have shown that COVID-19 has been a trending topic in recent years. When the conceptual structure map was examined, scale, reliability, and validity were included in a separate dimension. When the network map was examined, item response theory, differential item functioning, psychometrics, assessment, measurement, Rasch model, measurement invariance, item response theory (IRT), classical test theory, computerized adaptive testing, graded response model, confirmatory factor analysis, validation, personality, partial credit model, psychometric properties, missing data, scale development, simulation, dimensionality, nonparametric response item theory, construct validity, validity concepts, multidimensionality, screening, gender, generalized partial credit model, person fit, model fit keywords were in one cluster. The keyword network map had 50 concepts and 6 clusters. Word analysis conducted in Aksu and Güzeller's (2019, p. 61) research found that the most repeated words were "item response theory, classical test theory, model, validating, reliability, validity, and Rasch model," with overlapping this research.

There were some limitations to the research. The study had five criteria. The first criterion was related to the research topic. It is a criterion to have the expressions "item response theory" (IRT) or "item response modeling," or "item response model" in the abstract. Terms in the abstract could be searched by adding synonyms (for example, latent trait models). Only articles were included in the research. Publications such as book chapters and conference papers could be included in bibliometric analysis. Publications related to education and psychology were discussed in the research. There were also publications in the field of health-related to IRT. Research could be carried out by including other research areas. The research focused on publications between 2001 and 2021 and did not include articles published in other years. Similar research can be conducted by expanding the year range. Searching all articles in SSCI or ESCI or SCI-E or A&HCI was one of the inclusion criteria. However, there were many IRT related articles searched in other sources. More extensive research can be done on different datasets (e.g., Scopus). Finally, it should be noted that bibliometric research will guide researchers and reveal gaps in the field.

Declaration

Author Contribution: SB wrote all sections including "abstract, introduction, method, findings, conclusion and discussion". SB had roles in the conceptualization, resources, data analysis, reporting, drafting, reviewing, and editing.

Conflict of Interest: There is no conflict of interest.

Ethical Approval: Ethical rules were followed in this research. Ethical approval is not required, because data from WoS was used in this research.

References

Aksu, G., & Güzeller, C. O. (2019). Analysis of scientific studies on item response theory by bibliometric analysis method. *International Journal of Progressive Education*, *15*(2), 44-64. https://doi.org/10.29329/ijpe.2019.189.4

Andrés, A. (2009). *Measuring academic research: How to undertake a bibliometric study*. Elsevier.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Bailón-Moreno, R., Jurado-Alameda, E., Ruiz-Baños, R., & Courtial, J. P. (2005). Analysis of the field of physical chemistry of surfactants with the Unified Scienctometric Model. Fit of relational and activity indicators. *Scientometrics*, *63*(2), 259-276. https://doi.org/10.1007/s11192-005-0212-4

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21-33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Chen, C. (2006). CiteSpaceII: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609-1630. https://doi.org/10.1002/asi.22688

Demir, S. B. (2018). Predatory journals: Who publishes in them and why? *Journal of Informetrics*, *12*(4), 1296-1311. https://doi.org/10.1016/j.joi.2018.10.008

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W.M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Durieux, V., & Gevenois, P. A. (2010). Bibliometric indicators: Quality measurements of scientific publication. *Radiology*, *255*(2), 342-351. https://doi.org/10.1148/radiol.09090626

Egghe, L. (2005). Expansion of the field of informetrics: Origins and consequences. *Information Processing and Management, 41*(6), 1311–1316. https://doi.org/10.1016/j.ipm.2005.03.011

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Lawrence Erlbaum.

Eysenck, H. J. (1973). *The measurement of intelligence*. Medical & Technical Publishing Co.

Garfield, E. (1955). Citation indexes for science. *Science*, *122*(3159), 108-111. https://doi.org/10.1126/science.122.3159.108

Garfield, E. (1964). "Science Citation Index"-a new dimension in indexing. *Science*, *144*(3619), 649-654. https://doi.org/10.1126/science.144.3619.649

Gómez Benito, J., Hidalgo Montesinos, M. D., Guilera Ferré, G., & Moreno Torrente, M. (2005). A bibliometric study of differential item functioning. *Scientometrics*, *64*(1), 3-16. https://doi.org/10.1007/s11192-005-0234-y

Grauwin, S., & Sperano, I. (2018). Bibliomaps-a software to create web-based interactive maps of science: The case of UX map. *Proceedings of the Association for Information Science and Technology*, *55*(1), 815-816. https://doi.org/10.1002/pra2.2018.14505501129

Gülmez, D., Özteke, İ., & Gümüş, S. (2020). Overview of educational research from turkey published in international journals: A bibliometric analysis. *Education & Science*, *46*(206), 213-239. http://dx.doi.org/10.15390/EB.2020.9317

Gümüş, S., Bellibaş, M. Ş., Gümüş, E., & Hallinger, P. (2019). Science mapping research on educational leadership and management in Turkey: A bibliometric review of international publications. *School Leadership & Management, 40*(1), 1-22. https://doi.org/10.1080/13632434.2019.1578737

Hallinger, P., & Hammad, W. (2019). Knowledge production on educational leadership and management in Arab societies: A systematic review of research. *Educational Management Administration & Leadership, 47*(1), 20-36. https://doi.org/10.1177/1741143217717280

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

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management*, *45*(2), 175-194. https://doi.org/10.1177/0312896219877678

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Mingerne*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

National Center for Education Statistics (NCES). (2020). *Number of degree-granting postsecondary institutions and enrollment in these institutions, by enrollment size, control, and classification of institution: Fall 2020*. https://nces.ed.gov/programs/digest/d21/tables/dt21_317.40.asp

National Science Foundation (NSF). (2007). *Asia's rising science and technology strength: Comparative indicators for Asia, the European Union, and the United States*. http://www.nsf.gov/statistics/nsf07319/pdf/nsf07319.pdf

Osareh, F. (1996). Bibliometrics, citation analysis and co-citation analysis: A review of literature I. *Libri, 46(*3), 149-158. https://doi.org/10.1515/libr.1996.46.3.149

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use Bibexcel for various types of bibliometric analysis. *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday*, *5*, 9-24. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-25636

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, *25*(4), 348-349. https://doi.org/10.1108/eb026482

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Rey-Rocha, J., Martín-Sempere, M., & Garzon, B. (2002). Research productivity of scientists in consolidated vs. non-consolidated teams: The case of Spanish university geologists. *Scientometrics*, *55*(1), 137-156. https://doi.org/10.1023/a:1016059222182

Sci^{2}Team. (2009). *Science of Science (Sci^2) Tool*. Indiana University and SciTech Strategies. http://sci.slis.indiana.edu

Sengupta, I. N. (1992). Bibliometrics, informetrics, scientometrics and librametrics: An overview. *Libri, 42*,75–98. https://doi.org/10.1515/libr.1992.42.2.75

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics, 8*(4), 802-823. https://doi.org/10.1016/j.joi.2014.07.006

van Eck, N. J., & Waltman, L. (2010). Software survey: VoSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

van Raan, A. F. (2004). Measuring science. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 19-50). Wolters Kluwer. Springer.

Venables, W. N., Smith, D. M., & R Development Core Team. (2021). *An introduction to R*. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Ware M., & Mabe, M. (2015). *The STM report. An overview of scientific and scholarly journal publishing.* The Hague: International Association of Scientific, Technical and Medical Publishers. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1008&context=scholcom

Yurtçu, M., & Güzeller, C. (2021). Bibliometric analysis of articles on computerized adaptive testing. *Participatory Educational Research, 8*(4), 426-438. https://doi.org/10.17275/per.21.98.8.4

Zupic, I., & Cater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods, 18*(3), 429-472. https://doi.org/10.1177/1094428114562629

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 194-211

Latent Growth Modeling of Item Process Data Derived From Eye-tracking Technology: An Experimental Study Investigating Reading Behavior of Examinees When Answering A Multiple-Choice Test Item

Ergün Cihat ÇORBACI* Nilüfer KAHRAMAN**

Abstract

This study illustrates how eye-tracking data can be translated to "item process data" for multiple-choice test items to study the relationship between subjects' item responses and choice reading behavior. Several modes of analysis were used to test the hypothesized added value of using process data to identify choice reading patterns of subjects. In addition to the cross-sectional analises of agggregate measurements derived from the time series eye tracking data, Latent Growth Curve Model Analises were undertaken to test if the the shape of change observed in the sequential choice reading patterns differed for subjects depending on their responses to the item being correct or incorrect. Application data were from an experimental study and included seventy-one subjects' responses to two multiple-choice test items measuring reading comprehension ability in English as a second language. Analyses were carried out for one item at a time. For each item, first, each subject's recorded eye movements were coded into a set of Area of Interests (AOIs), segmenting the lines in the stem and the individual choices. Next, each subject's fixation times on the AOIs were time stamped into seconds, indicating when and in what order each subject's gaze had fixated on each AOI until a choice was marked as the correct answer, which ended the item encounter. A set of nested Latent Growth Curve models were considered for the choice-related AOIs to deliniate if distinct choice-process sequences were evident for correct and incorrect respoders. Model fit indices, random intercepts, slopes, and residuals were computed using the mean log fixation times over item encounter time. The results show that the LGM with the best model fit indicies, for both items, was the quadratic model using response variable as a covariate. Albeit limited due to the two-item – seventy-one subjects experimental setting of the study, the findings are promising and show that utilizing item-level process data can be very useful for defining distinct choice processing (task-oriented reading) patterns of examinees. Over all, the results warrant further study of choice derived AOIs using longitudinal statistical models. It is argued that, the screening methodology desribed in this study can be a useful tool to investigate speededness, distractor functioning, or even to flag subjects with irregular choice processing behavior, such as providing a direct mark on a choice, without any significant reading activity on any of the choices presented (i.e., whether cheating might have occurred.)

Keywords: Latent Growth Curve Modeling, eye-tracking, reading ability in English, multiple-choice items.

Introduction

The eye-tracking technology has been widely used for investigating how individuals read words or sentences and whether tracking their reading behavior while reading can be helpful to understand the cognitive processes functioning (Rayner, 1998). However, the use of time series eye-tracking data to improve educational assessment settings, where examinees are to answer questions given a text that is specifically constructed to measure reading comprehension ability (i.e., task-oriented reading), has been neglected to a great extent, which can potentially support and enrich reliability and validity studies

* Research Assistant, Gazi University, Faculty of Education, Ankara-Turkey, e.cihat.corbaci@gmail.com, ORCID ID: 0000-0002-7874-956X
** Professor Doctor, Gazi University, Faculty of Education, Ankara-Turkey, nkahraman@gazi.edu.tr, ORCID ID: 0000-0003-2523-0155

To cite this article:
Çorbacı, E. C., & Kahraman, N. (2022). Latent growth modeling of item process data derived from eye-tracking technology: An experimental study investigating reading behavior of examinees when answering a multiple-choice test item. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 194-211. https://doi.org/10.21031/epod.1107597

Received: 22.04.2022
Accepted: 25.08.2022

focusing on various measurement processes (Solheim & Uppstad, 2011). As the Standards for Educational and Psychological Testing (American Educational Research Association, et al., 2014) states "The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations". Eye movement data can be useful in the study of how examinees process the choices of a multiple-choice test item, that is, before providing a response (True/false). Identifying response processes patterns through measures derived from process data such as gaze-movements on a screen, although indirectly, could reveal item and domain specific features of test scores and uses. Moreover, different test-taker groups such as competent and incompetent test-takers may respond to a multiple-choice question using different patterns, which can be used to verify or falsify a proposed interpretation (Kane & Mislevy, 2017).

There are relatively few studies, in the literature, that underline the importance of investigating assessment-related aspects of such inquiries. Paulson and Henry (2002), for example, used eye-tracking movements to scrutinize claims (measure the reading comprehension process) asserted by the publisher on a reading assessment (Degrees of Reading Power, DRP) and investigated the reading processes of test-takers while taking DRP. They used a modified cloze setup of DRP that was intended to measure the process of reading by responding to the comprehension questions at the end of the passage. Tai et al. (2006) also used eye-tracking movements to investigate problem-solving behaviors within a group of subjects in three different disciplines while solving standardized multiple-choice questions. They analyzed the location of eye-gaze fixation, duration of fixation, scan paths, and duration between fixations as well as correct responses and latent response times, which consist of both quantitative and qualitative data analysis. Solheim and Uppstad (2011) also used eye-tracking to investigate problem-solving behaviors using a stimulus text comprised of both a verbal text and an illustration. They related correct-answer scores to gaze movement patterns arguing that subjects' gaze movement behavior revealed subjects' reading behavior. Tsai et al. (2012) examined students' visual attention when solving a multiple-choice science problem using an eye-tracker. They divided students into two groups: the high-score group and the low-score group, and unlike other studies, they investigated choices (distractors and the correct choice) in the multiple-choice questions. In addition, they stated that students paid more attention to the options they chose and to relevant areas and paid little attention to the irrelevant areas. Yaneva et al. (2022) demonstrated how to use multiple-choice questions to collect evidence for validity argument. They investigated how the presence of options in the multiple-choice question affect the response behavior of the students, what areas of item they viewed first and whether the options were processed in the same way, discussing validity inferences. Overall, considering the studies in the literature, it can be seen that measurement-based approaches used in education include traditional reading and task-oriented readings. However, it is critical that there is a methodological perspective that can assist in demonstrating the validity and reliability of such approaches.

To this end, this paper proposes and illustrates a two-stage methodology highlighting the necessity of an in-synch multi-stage data processing approach when integrating eye-tracking technology-derived (response-related) data into the conventional psychometric analysis that most often uses response data alone. Formulated to place a special emphasis on a data screening stage to be carried out prior to the actual data analysis stage, the proposed methodology is illustrated using real response data collected for a couple of multiple-choice items from a test measuring college students' reading comprehension ability in English as a Foreign Language. The application presented helps demonstrate that it pays off to investigate the inner-connectedness of research questions to the information available in the eye-tracking (device recordings, i.e., gaze durations and movements coded in milliseconds) and the conventional response data (i.e., given a question, markings for the correct choice given choices from A to E), as a priory to the final data analysis stage. It is argued that the stronger relationship is between the measurement variables created out of eye-tracking data and the desired interpretations, the easier it will be to make inferences about the findings for the integrated response process data (eye-tracking recordings + item responses). Underlining the importance of using a psychometric perspective when analyzing eye-tracking-aided item response data, the ultimate purpose of this study is to provide several modeling strategies that can help researchers capture construct-related information that might be available in eye-tracking data and to test the meaningfulness of its added value.

Method

Data

Application data were from a test experiment using items from a multiple-choice reading comprehension test in English. The data set included both the conventional item response data, i.e., the choice marked by the subjects as the correct answer (0/1), and the eye-tracked process data, i.e., fixation durations and sequences over item-encounter time given the area of interests, AOIs, for two separate multiple-choice test items. The test experiment, its subjects and how the AOIs were defined are described in the following text.

Test Experiment and Subjects

The eye-tracking apparatus used for data collection was the Tobii TX300 screen-based eye tracker, which performed binocular tracking with a sampling frequency 300Hz (Dell Desktop Computer, Intel Core i7-4790 @3.60 GHz, 16,0 GB of RAM). Seventy-one subjects took part in the test experiment. The subjects were non-native speakers of English. Each examinee took the exam individually and responded to test items appearing on a computer screen one at a time, in the same order. The items were from the released items of the Foreign (English) Language Exam that was administered by the Student Selection and Placement Center (OSYM) in 2018. This exam is held twice a year and consists of 80 multiple-choice items that includes different contents such as vocabulary, grammar, translation etc. to evaluate foreing language skills. In addition, this test evaluates only reading comprehension skills, which means there is no questions related to other skills such as listening, speaking and writing.

Processing the Eye-Tracking Data

In the test experiment, the layout on the test screen placed the stem part of the item on the left side of the screen, while the choices were placed on the right side. Figure 1 shows the ten areas of interest (AOI) used for Item 1: Direction, Line 1, Line 2, Line 3, Line 4, Choices A, B, C, D, and E. Figure 2 shows the nine AOIs used for Item 2: Direction, Line 1, Line 2, Line 3, Choices A, B, C, D, and E. As Figure 1 (Item 1-10 AOIs) and Figure 2 (Item 2 - 9 AOIs) illustrate, within the context of multiple-choice test items, unlike the AOIs for the choices A to E, the AOIs in the stem may differ drastically from one item to another due to item-specific features, such as the number of lines included or the number of words included in each line.

Figure 1

The Areas of Interest for the Multiple-Choice Item 1

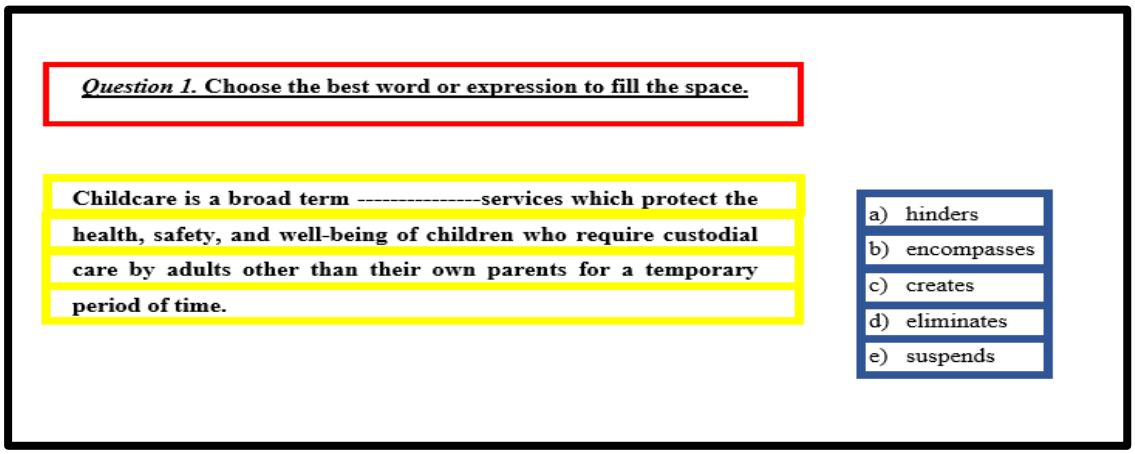
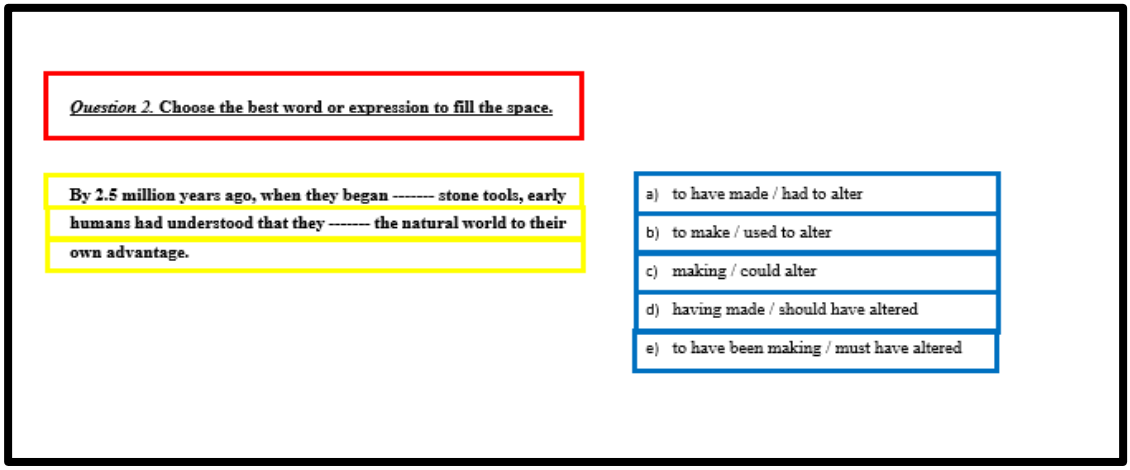


Figure 2

The Areas of Interest for a Multiple-Choice Item 2



In the data processing stage of this study, the AOIs presented in the corresponding figures above were used to binary code the eye-tracking data for each item. The data were exported using the defined item-AOIs: 300 rows of data per second (the data were collected with the sampling frequency of 300Hz, where a piece of new information was collected in each 3.34 ms). Considering that each student spent approximately one minute answering an item, a raw data file of about 18000 rows was obtained per item subject encounter.

Next, the data collected in the item experiments were screened to detect and remove corrupt, irrelevant, and inaccurate recordings from the dataset composed of (1) response (correct-incorrect coding, 1-0), and fixation durations over (2) lines within item stems (text part of the items) and (3) choices A to E. One of the fundamental problems in eye-tracking data is the desynchronization between gaze and stimuli resulting from poor calibration or visual impairment. To overcome this problem, as a first step, subjects who never looked at the relevant areas and therefore did not have any eye movements in these areas were determined through a careful analysis of the corresponding time graphs. In addition, the subjects with gaze-stimulus mismatches were identified by examining each examinee's scan paths one by one. For example, Figure 3 below shows the scan path, and Figure 4 shows the heatmap of a subject's encounter with an item illustrating how the desynchronization between the subject's gaze and stimuli might occur, rendering the validity of the information available in the resulting eye-tracking process data for this subject.

The dots, in Figure 3, show where the subject is looking (fixations), and the numbers in these dots show the order in which an individual looks/fixates, and the lines connecting these dots show the transitions (saccades). As one can see, the subject looked at the stem of the question, but the gaze movements were lower. Similarly, the subject actually looked at choice A, but did not seem to look at it because their eye movements were shifted downwards.

Figure 4 shows where the subject is looking and focusing. As in the scan path, the subject looked at the stem of the question, but the gaze movements were lower. Similarly, the subject actually looked at choice A, but did not seem to look at it because their eye movements were shifted downwards. In general, the subject's reading movements shifted below the lines and choices. Process data collected for each item-subject encounter were screened to spot desynchronizations.

Figure 3

The Scan Path for Item 1 for Subject 12011 Illustrating Desynchronization

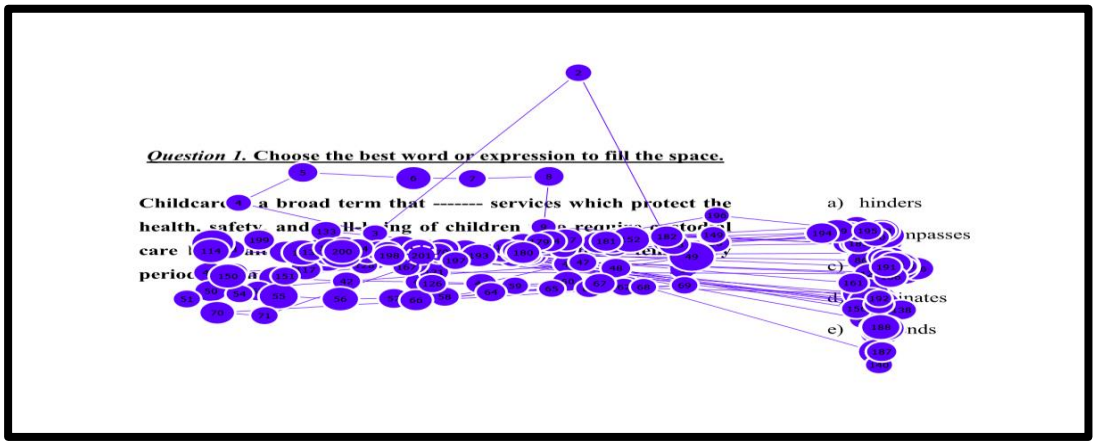


Figure 4

The Heatmap for Item 1 for Subject 12011 Illustrating Desynchronization

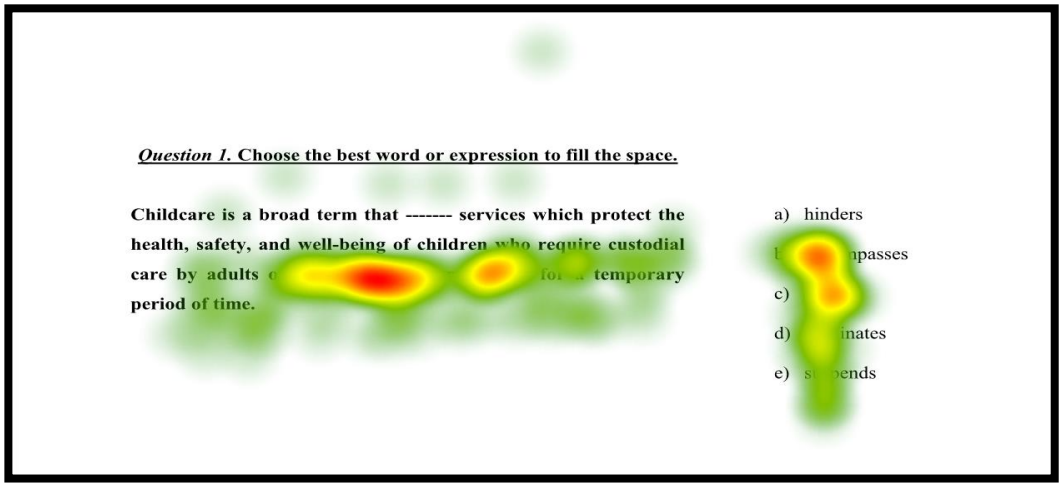


Figure 5 illustrates another subject's eye movements for Item 1. Unlike the Figure 3, Figure 5 shows that this subject's gaze movements and the lines are overlapping, which is desired for the data quality. Figure 6 illustrates the heatmap for the second subject for Item 1. As one can see, the subject's gaze movements and the lines are overlapping, and the subject focuses on the first line and choices; however, there is no downward or upward movement beyond the lines.

Figure 5

The Scan Path for Item 1 for Subject 39770 Illustrating Synchronization

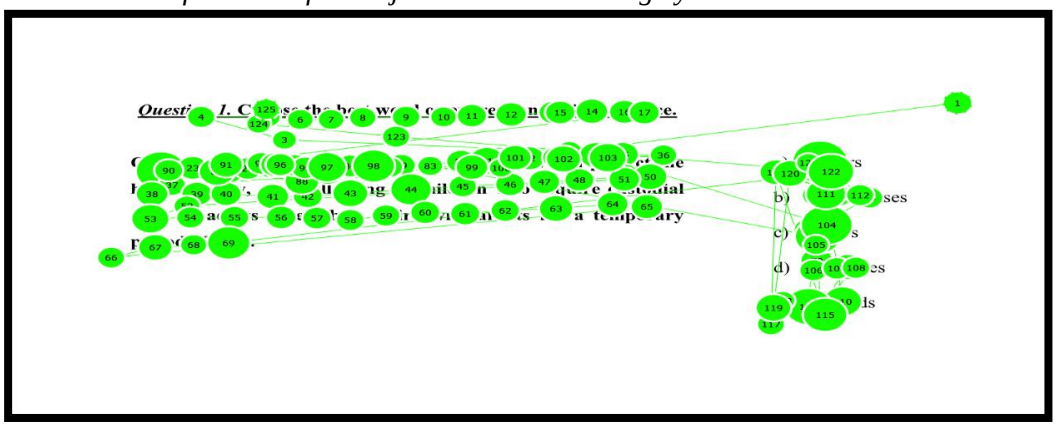
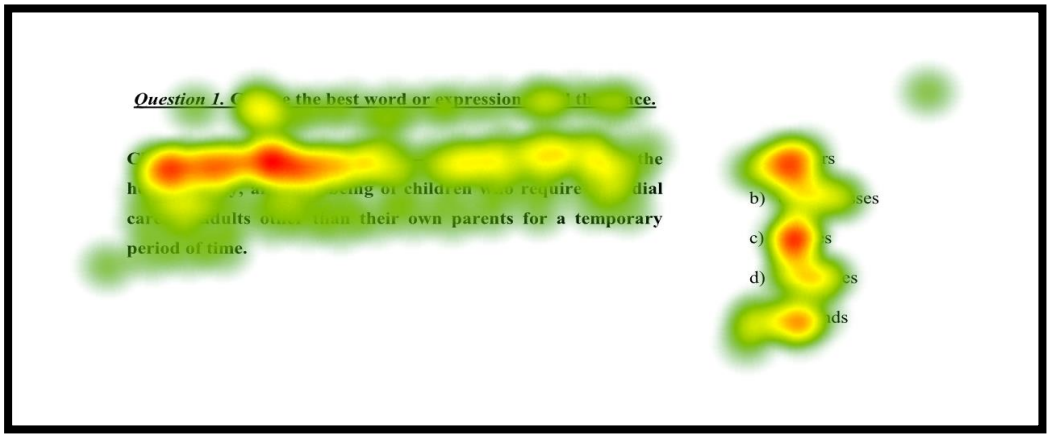


Figure 6

The Heat Map for Item 1 for Subject 39770 Illustrating Synchronization



In this study, utilized eye-tracking metrics were "fixation-related" recordings of each subject-item encounter. These metrics can be processed to obtain the number of fixations, fixation duration, total fixation duration, average fixation duration, fixation rate, regressive fixation, number of saccade, saccadic duration (see Holmqvist et al., 2011, for an elaborative list of measures). In this study, only fixation-related measures such as total gaze duration, total fixation duration, mean fixation duration and the mean fixation that was used to refer to the means that were computed from the choice sequence process data" were used.

Table 1 gives summary statistics calculated for the areas of interest (direction, line, and choices) determined for item 1. This table was created by aggregating time-stamped data (it can be seen in the Figures 3 and 5 (scan paths), which aggregated graphical version of these miliseconds into fixations and saccades, showing the hit sequence and fixation densities of eye-movements) provided by eye-tracking device, which collects data every 3.4 ms. The generated data set in Table 1 lists numerical values that can be thought of as their quantitative counterparts needed for statistical analysis and model fitting to describe and test the hypotheses about whether there was a within person change over the choices presented A to E, if so, which parameters would be needed to predict the shape of change implied by the data.

Table 1

Descriptive Statistics for Item 1 Based on Fixation Measures for Subject 76798

Areas of Interest	Hit Sequence	Time to First Fixation	Total Fixation Duration (ms)	Total Dwell Duration	Total Fixation Count
Direction	2	0,84	1,77	2,1	10
Line 1	1	0,4	10,91	12,25	59
Line 2	8	12,25	2,81	3,44	17
Line 3	9	18,67	1,93	2,28	9
Line 4	10	21,72	0,22	0,23	2
Choice A	3	5,56	3,96	4,66	17
Choice B	4	6,23	3,58	3,69	16
Choice C	6	7,64	2,58	2,78	12
Choice D	5	7,48	3,95	4,34	18
Choice E	7	10,19	2,41	2,8	11

Analysis of Eye-Tracking Process Data

As stated in the previous section, the eye-tracking device provides process data sampled at regular intervals. The process data in this research were collected each 3.4 ms using the eye-tracking technology and were restructured for the cross-sectional analysis and the latent modeling on each item. As a first step, summary statistics were computed for Item 1 and Item 2 by using total gaze duration and mean log fixation duration for the direction, stem and the choices. Then, it was tested whether these durations were differed for groups who responded correctly/incorrectly to the item. After that, for the latent modeling, mean log fixation times were calculated using time-stamped data by averaging each time the test-taker reads an area of interest (choice A, B, C, D and E), from starting to read an item until a response was provided. In this context, it was hypothesized that the default processing order was the given presentation order of item segments from A to E (within-person variance). It was hypothesized if this were the case for the sample, along with threshold and slope variances showing between-examinee variances.

Latent Growth Curve Modeling For Eye-Tracking Response Data

Although many techniques have been used to analyze eye-tracking data, the usefulness of Latent Growth Curve Modeling remains relatively unexplored. It is, therefore, worthwhile to explore what model formulations can offer for process data analytics and their connections to the latent structure intended to be measured. Permiting both within-person change over time and between-person variability, Latent Growth Curve Models are used to answer questions such as: "What is the shape of the mean trend over time?, Does the initial level predict the rate of change? Do two or more groups differ in their trajectories? etc." (Duncan et al., 2013; Muthen, 2001; Preacher at. al., 2008; Willett & Sayer, 1994).

In this study, LGCM applications were utilized to model change over response choices (AOIs for A to E) for one item at a time. For this purpose, a series of nested latent growth curve models were fit to item-level data to estimate (a) choice mean log fixation time trajectories and (b) variability around the initial status and overall rate of change. Relying on the conventional assumption that subjects read the choices in given presentation order, a linear model was considered first. The linear latent growth model curve model was estimated by Equation 1:

Level 1 within-subject model: $$y_{it} = \eta_{0i} + \eta_i a_{it} + \varepsilon_{it},$$

Level 2 between-subject model: $$\eta_{0i} = \beta_0 + \zeta_{0i}, \quad (1)$$

$$\eta_{1i} = \beta_1 + \zeta_{1i},$$

where *t* represents the choice set coded 0 to 4, *i* represents the subject (*I=1,2,…, N*), *y* represents the mean log fixation time influenced by the random effects η_{0i} and η_{1i}. The intercept η_{0i} describes a subject's initial mean log fixation time when reading the choices. The linear term η_{1i} describes the rate of change in mean log fixation time during the reading of the choices.

The quadratic latent growth curve model was considered next. The model was estimated by Equation 2:

Level 1 within-subject model: $$y_{it} = \eta_{0i} + \eta_{1i} a_{it} + \eta_{2i} a_{it}^2 + \varepsilon_{it},$$

Level 2 between-subject model: $$\eta_{0i} = \beta_0 + \zeta_{0i}, \quad (2)$$

$$\eta_{1i} = \beta_1 + \zeta_{1i},$$

$$\eta_{2i} = \beta_2 + \zeta_{2i},$$

where η_{2i} represents the quadratic term, and it describes the rate of acceleration in the rate of change over the course of reading choices.

The quadratic latent growth curve model with the item response (correct/incorrect) as a time-invariant covariate was considered next. The model was estimated by Equation 3:

Level 1 within-subject model: $$y_{it} = \eta_{0i} + \eta_{1i} a_{it} + \eta_{2i} a_{it}^2 + \varepsilon_{it},$$

Level 2 between-subject model: $$\eta_{0i} = \beta_0 + \gamma_0 X_i + \zeta_{0i}, \quad (3)$$

$$\eta_{1i} = \beta_1 + \gamma_1 X_i + \zeta_{1i},$$

$$\eta_{2i} = \beta_2 + \gamma_2 X_i + \zeta_{2i},$$

Where γ_0, γ_1 and γ_2 *y* represents the parameters for the association between the covariate, X_i, and each latent growth term.

Log transformations were used in model estimations to normalize mean fixation time distributions. The models were estimated with Mplus using the Maximum Likelihood estimator. Consistent with recommended practices (see Hu & Bentler, 1999; for a detailed discussion), more than one fit index was used in evaluating model fit. The models were compared using (a) the Comparative Fit Index (CFI; Bentler, 1990), and (b) the Tucker-Lewis Index (TLI; Bentler & Bonnet, 1980), where values range from zero to one and the values greater than 0.95 may be interpreted as an acceptable fit; (c) the root Mean Square Error of Approximation (RMSEA; Steiger, 1990), where values smaller than 0.05 indicate a good model fit (Browne & Cudeck, 1993); (d) Standardized Root Mean Square Residual (SRMR; Bentler, 1995), where values should be less than 0.05 for a good fit (Hu & Bentler, 1998), (e) Akaike Information Criterion (AIC; Akaike, 1974) the Bayesian Information Criterion (BIC; Schwarz, 1978), where smaller values indicate a better fit and are often used to test goodness-of-fit for a full model in comparison with a reduced one.

Results

For the data obtained from the eye tracking device to be valid and reliable, the correct positioning of the eye movements on the stimuli is one of the essential points to be considered. Even though the subjects in the study stated that they did not have any visual impairment and passed the calibration test, it was found that the desired quality of stimulus-eye movement harmony was somewhat flactuating for some subjects while steadily unusable for some others. A careful screening of the scanmaps and heat maps

revealed that, out of 71 subjects in the sample, eye-tracking recordings of 11 subjects were unusable for item 1, while only 6 of these subjects' eye-tracking recordings were also unusable for item 2.

Table 2 shows the summary data (gaze even duration and the mean fixation duration) calculated for the areas of interest (from question stem to lines) for all subjects in item 1 and item 2. The table show that the total-time subjects spent on Item 1 (gaze event duration) varied between 8,23 seconds and 74,76 seconds, and the mean duration was 36,07 seconds. There was a considerable variation in how much time the subjects spent answering the item. Similarly, there was an apparent variation in how much time the subjects spent on each choice and each line considering the standard deviations. The minimum time spent on some areas of interest, such as stem, choice A, and line 2, was zero, which means that at least one subject responded to the question without looking at these areas. Among the choice, the true choice (choice B) had the maximum time spent, and of the lines, the first line had the maximum time spent.

The total-time subjects spent on item 2 varied between 15,35 seconds and 97,42 seconds, and the mean fixation duration was 39,82 seconds. Similarly, there was an apparent variation in the duration the subjects spent responding to item 2, each choice, and each line, considering their standard deviations. The minimum time spent on some areas of interest, such as stem, choice B, choice E, and line 2, was zero, which means that at least one subject responded to the question without looking at these areas. Unlike item 1, the true choice (choice C) didn't have the maximum time spent, and of the lines, the first line had the maximum time spent.

Considering the right/wrong answer, which is one of the most basic features of a multiple-choice items, subjects were grouped into two groups (subjects who responded correctly vs. incorrectly) to explore whether the subjects who gave correct answers allocated their attention to different parts of the items differently from the subject who gave an incorrect answer.

Table 2

Descriptive Statistics for Item 1 and Item 2 Based on Gaze Event Duration and Mean Fixation Duration for All Subjects

			Fixation Durations								
Item 1											
	Total Gaze Dur.	**Stem**	**Line 1**	**Line 2**	**Line 3**	**Line 4**	**Cho. A**	**Cho. B***	**Cho. C**	**Cho. D**	**Cho. E**
Mean	36,07	1,93	8,13	5,58	2,65	0,68	1,79	2,8	2,64	2,18	1,69
Std. Deviation	16,01	1,80	5,08	3,95	2,01	0,69	1,44	1,4	1,82	1,73	1,64
Median	35,08	1,65	7,5	4,47	2,19	0,51	1,3	2,57	2,24	1,48	1,08
Minimum	8,23	0,00	0,1	0,18	0	0	0	0,62	0,43	0,16	0,19
Maximum	74,76	9,87	22,77	17,43	8,89	3,79	6,6	6,37	8,86	6,98	7,71
Item 2											
	Total Gaze Dur.	**Stem**	**Line 1**	**Line 2**	**Line 3**	**Line 4**	**Cho. A**	**Cho. B**	**Cho. C***	**Cho. D**	**Cho. E**
Mean	39,82	1,51	9,46	7,73	0,86	N/A	3,39	3,58	3,29	2,28	1,53
Std. Deviation	16,50	1,38	4,82	4,13	0,73	N/A	2,58	2,94	1,89	1,86	1,54
Median	39,99	1,35	8,87	6,88	0,73	N/A	2,77	3,04	2,89	1,63	1,08
Minimum	15,35	0,00	1,38	1,91	0	N/A	0,28	0	0,55	0,26	0
Maximum	97,42	6,12	22,48	19,56	4,36	N/A	10,33	17,04	10,4	11,44	7,58

* the correct answer

Table 3 illustrates the descriptive statistics for the two groups (Item 1). The *First Group* consisted of those who did not score any point on Item 1 (n = 36), and the *Second Group* consisted of those who scored one point on Item 1 (n = 24). The first group had higher *mean total gaze duration* and *mean fixation durations* for each variable except for *choice B*, the correct choice, than the second group. According to the Mann-Whitney U test results, there was a statistically significant difference between the first and second groups regarding total gaze duration, choice A, C, D, E, and Lines 1,2,3,4 ($p < .05$).

Table 4 also illustrates the decriptive statistics for two groups (Item 2). As in item 1, the *First Group* consisted of those who did not score any point on Item 2 (n = 40), and the *Second Group* consisted of those who scored one point on Item 2 (n = 25). The first group had higher *mean total gaze duration* and *mean fixation durations* for each variable except for *choice C*, the correct choice, than the second group. According to the Mann-Whitney U test results, there was a statistically significant difference between the first and second groups regarding total gaze duration, choices A, B, D, E, and Lines 1 and 2 ($p < .05$).

Table 3

Descriptive Statistics for Item 1 Based on Mean Fixation Duration in Terms of Correct-Incorrect Response Groups

Item 1	Fixation Durations										
	Total Gaze Dur.**	**Inst.**	**Line 1****	**Line 2****	**Line 3****	**Line 4****	**Opt. A****	**Opt. B**	**Opt. C****	**Opt. D****	**Opt. E****
Subjects who answered the item incorrectly											
Mean	42,39	2,17	9,4	6,89	3,09	0,83	2,14	2,78	3,44	2,85	2,17
Std. Deviation	13,89	1,92	5,05	4,31	2,06	0,75	1,57	1,46	1,83	1,88	1,75
Median	39,09	1,94	9,47	5,49	2,6	0,59	1,55	2,63	3,58	2,17	1,61
Minimum	19,05	0,00	0,31	1,46	0	0	0,16	0,62	0,88	0,41	0,21
Maximum	74,76	9,87	22,77	17,43	8,89	3,79	6,6	6,3	8,86	6,98	7,71
Subjects who answered the item correctly											
	Total Gaze Dur.**	**Inst.**	**Line 1**	**Line 2**	**Line 3**	**Line 4**	**Opt. A**	**Opt. B**	**Opt. C**	**Opt. D**	**Opt. E**
Mean	26,58	1,56	6,24	3,6	1,98	0,46	1,26	2,83	1,45	1,17	0,96
Median	21,54	1,39	5,05	3,14	1,76	0,28	0,91	2,49	1,18	1,1	0,41
Std. Deviation	14,39	1,56	4,6	2,21	1,76	0,53	1,05	1,34	0,98	0,71	1,16
Minimum	8,23	0,00	0,1	0,18	0	0	0	1,15	0,43	0,16	0,19
Maximum	59,15	7,02	17,84	9,45	7,17	1,75	4,11	6,37	4,68	3,03	4,96

* the correct answer = Choice B, ** $p < 0.05$

Table 4

Descriptive Statistics for Item 2 Based on Mean Fixation Duration for Terms of Correct-Incorrect Response Groups

Item 2	Fixation Durations									
	Total Gaze Dur.**	**Inst.**	**Line 1****	**Line 2****	**Line 3**	**Opt. A****	**Opt. B****	**Opt. C****	**Opt. D****	**Opt. E****
Subjects who answered the item incorrectly										
Mean	45,21	1,79	10,66	8,58	0,96	4,39	4,44	3	2,6	2,03
Std. Deviation	16,61	1,59	4,65	4,09	0,85	2,71	3,34	2,03	2	1,72
Median	42,30	1,37	9,19	7,46	0,74	3,87	3,66	2,76	1,89	1,58
Minimum	15,35	0,00	1,38	2,33	0	0,87	0	0,55	0,48	0,19
Maximum	97,42	6,12	22,48	19,56	4,36	10,33	17,04	10,4	11,44	7,58
Subjects who answered the item correctly										
	Total Gaze Dur.	**Inst.**	**Line 1**	**Line 2**	**Line 3**	**Opt. A**	**Opt. B**	**Opt. C**	**Opt. D**	**Opt. E**
Mean	31,35	1,06	7,56	6,39	0,69	1,82	2,23	3,74	1,78	0,76
Std. Deviation	12,52	0,83	4,56	3,92	0,45	1,25	1,36	1,58	1,54	0,75
Median	29,41	1,31	6,15	5,53	0,72	1,68	2,26	3,68	1,18	0,51
Minimum	16,32	0,00	1,57	1,91	0	0,28	0,16	1,71	0,26	0
Maximum	56,44	2,26	21,38	19,01	1,57	5,41	4,9	7,83	6,25	3,29

* the correct answer = Choice C, ** $p < 0.05$

Table 5 lists model fit statistics of item 1 for latent growth models identified in the research. Overall, the results suggest that the goodness-of-fit observed for the three models ranges from unacceptable to very good and that Model 2 with the quadratic term and Model 3 with the quadratic term and the response (0/1) as a covariate fit the data better than Model 1 with a linear term only. For Model 1, the fit statistics show that the model does not provide an adequate fit to the data as the RMSEA and SRMR far exceeds the acceptable fit range. For Model 2 and Model 3, the fit statistics suggest that these models have a very good fit to the data.

For comparing which one indicates a better model (Model 2 or Model 3), three information-based fit indices (Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample-size-adjusted BIC (SABIC)) were examined, where smaller values indicate a better model. These indices suggest that Model 3 fit the data better than its reduced counterparts. Besides, a likelihood ratio test was done to test the sufficiency of a complex model versus a smaller model. The significance value for this test ($9{,}737_{(2)}$) is 0.02, where one can accept a quadratic model with a covariate improved the model.

Table 6 shows random intercepts and slopes produced by different models. The results of Model 1 suggest that the subjects differed in two ways: in the estimates of their initial status (intercept) and their rates of change during the reading choices (slopes). Figure 7 plots intercepts and slope estimates for the five timing conditions (choices). This figure shows observed scores, Linear Model (Model 1) and Quadratic model (Model 2). It shows that the mean log fixation times of the subjects who took item 1 rose rapidly from Choice A to Choice B, while there were few increases from Choice B to Choice E.

Figure 8 illustrates the group-specific random coefficients (intercepts and slopes) produced by the Quadratic Model with Time-Invariant Covariate (Model 3) for the two response groups (correct/incorrect). Figure 8 also demonstrates that the correct-response group and incorrect response group differed in two respects: in the observed scores of their initial status (intercepts) and their rates of change during the fixation orders (slopes). While there is a downward trend in mean log fixation times after the correct choice B for the correct-choice group, there is an upward trend in the mean log fixation times from choice A to choice E for the incorrect-choice group.

Table 5

Model Fit Comparisons for Three Models for Item 1

	CFI	TLI	AIC	BIC	SABIC	RMSEA	SRMR
Model 1[a]	0, 96	0,96	272,89	293,88	262,38	0,14	0,13
Model 2[b]	0,99	0,99	266,77	296,09	252,06	0,07	0,05
Model 3[c]	0,99	0,99	253,26	288,87	235,40	0,06	0,05

[a]Linear Model,
[b]Quadratic Model,
[c]Quadratic Model with Time-Invariant Covariate

Table 6

Parameter Estimates for Three Models

		Intercept	Slope	Quadratic
Model 1	Mean	2.80 (0,00)	0.04 (0,02)	
	Variance	0.20 (0,00)	0.01 (0,04)	
Model 2	Mean	2.68 (0,00)	0.18 (0,00)	-0,03 (0,01)
	Variance	0.27 (0,00)	0.03 (0,40)	0,00 (0,23)
Model 3	Mean	2.82 (0.00)	0.20 (0.001	-0.03 (0.04)
	Variance	0.24 (0.00)	0.02 (0.46)	0.00 (0.25)

Figure 7

Observed Means and Expected Mean Log Fixation Times for Model 1 and Model 2

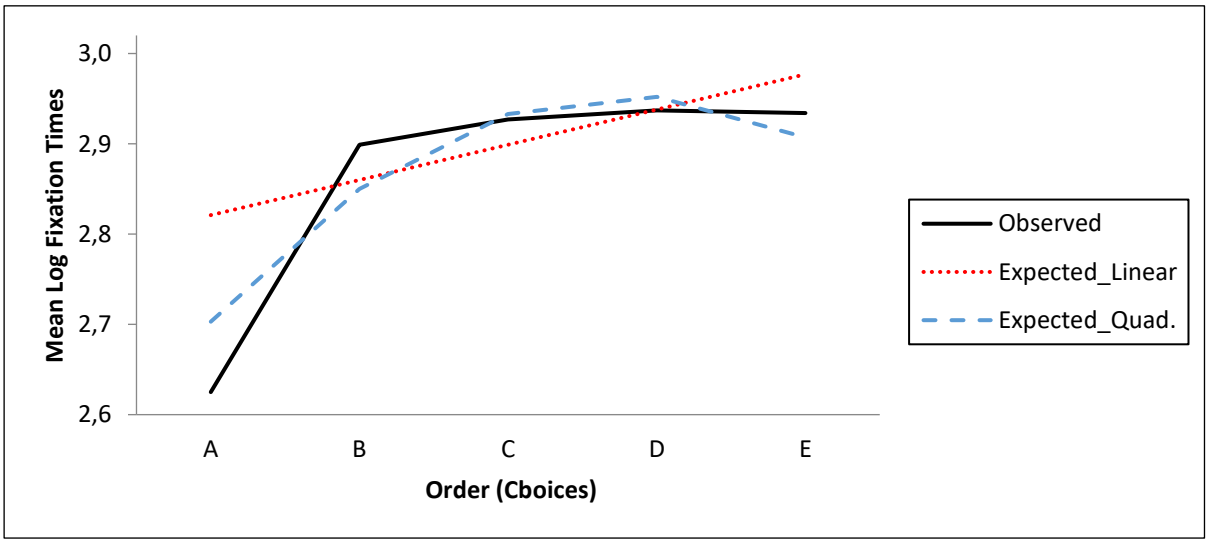


Figure 8

Estimated Mean Log Fixation Time for Correct-Response and Incorrect-Response Groups for Model 3

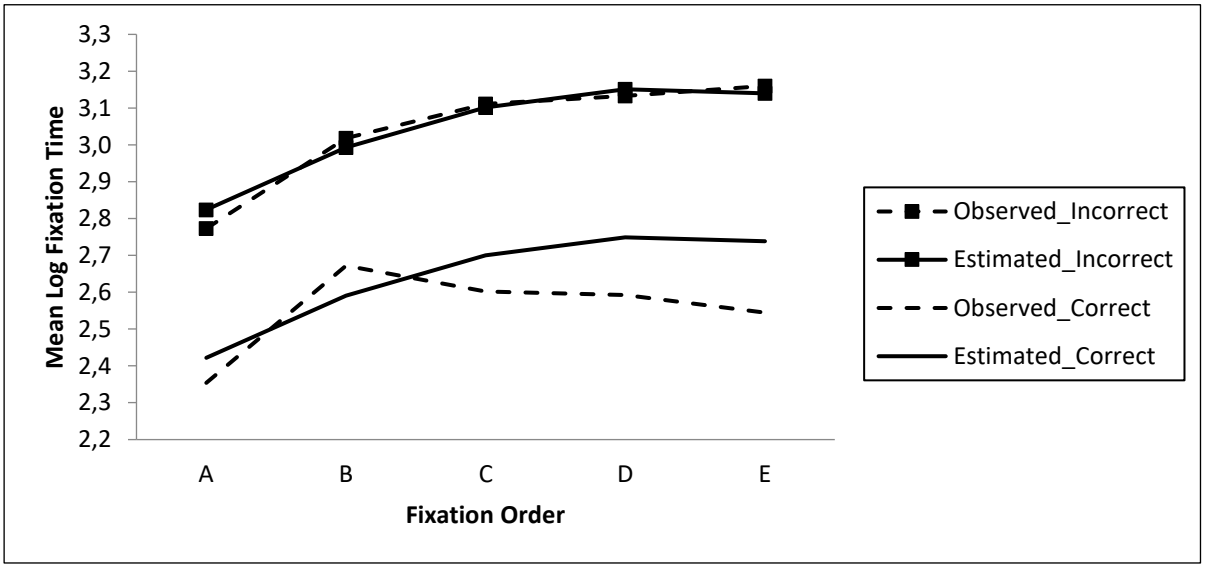


Table 8 lists model fit statistics of item 2 for latent growth models identified in the research. Overall, the results suggest that the goodness-of-fit observed for the three models ranges from unacceptable to very good and that Model 2 and Model 3 fit the data better than Model 1. For Model 1, the fit statistics show that the model does not provide an adequate fit to the data as the RMSEA and SRMR far exceeds the acceptable fit range. Although for Model 2 and 3, the statistics look similar, the fit statistics suggest that Model 3 has a very good fit to the data. To compare which one indicates a better model (Model 2 or Model 3), three information-based fit indices (Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and sample-size-adjusted BIC (SABIC)) were examined, where smaller values indicate a better model. These indices suggest that Model 3 fit the data better than its reduced counterparts. Besides, a likelihood ratio test was done to test the sufficiency of a complex model versus a smaller model. The significance value for this test ($6{,}214_{(2)}$) is 0.04, where one can accept a quadratic model with a covariate improved the model.

Table 9 shows random intercepts and slopes produced for Item 2 by different models. The results of Model 1 suggest that the subjects differed in two ways: in the estimates of their initial status (intercept) and their rates of change during the reading choices (slopes). Figure 9 illustrates that the mean log fixation times of the subjects who took item 2 rose rapidly from Choice A to Choice C, which is the correct choice, while there are few increases from Choice C to Choice E.

Table 8

Model Fit Comparisons for Three Models for Item 2

	CFI	TLI	AIC	BIC	SABIC	RMSEA	SRMR
Model 1[a]	0, 95	0,95	206,89	228,63	197,15	0,13	0,12
Model 2[b]	0,99	0,99	201,77	231,63	187,56	0,07	0,07
Model 3[c]	0,99	0,99	194,76	231,72	178,21	0,07	0,05

[a]Linear Model,
[b]Quadratic Model,
[c]Quadratic Model with Time-Invariant Covariate

Table 9

Parameter Estimates for Three Models for Item 2

		Intercept	Slope	Quadratic
Model 1	Mean	2.81 (0,00)	0.07 (0,00)	
	Variance	0.19 (0,00)	0.00 (0,02)	
Model 2	Mean	2.74 (0,00)	0.18 (0,00)	-0,02 (0,00)
	Variance	0.25 (0,00)	0.07 (0,02)	0,00 (0,02)
Model 3	Mean	2.91 (0.00)	0.09 (0.07)	-0.08 (0.45)
	Variance	0.19 (0.00)	0.06 (0.04)	-0.04 (0.03)

Figure 9

Observed Means and Expected Mean Log Fixation Times for Model 1 and 2

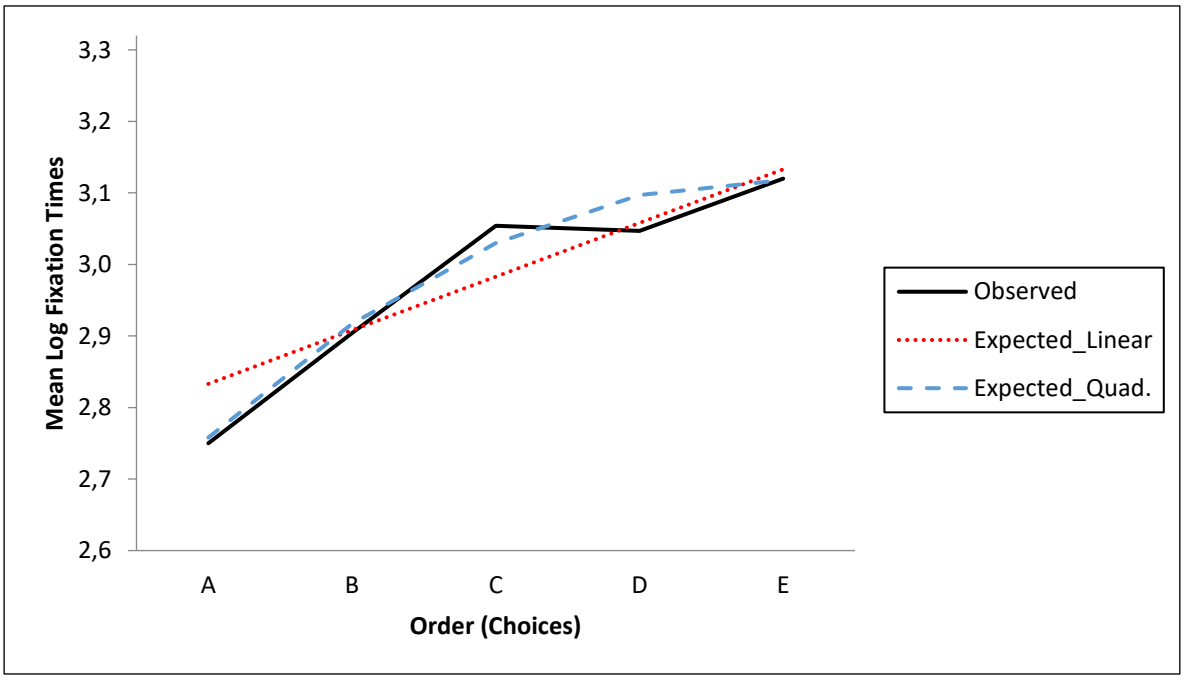


Figure 10

Model 3, Estimated Mean Log Fixation Time for Correct-Response and Incorrect-Response Groups

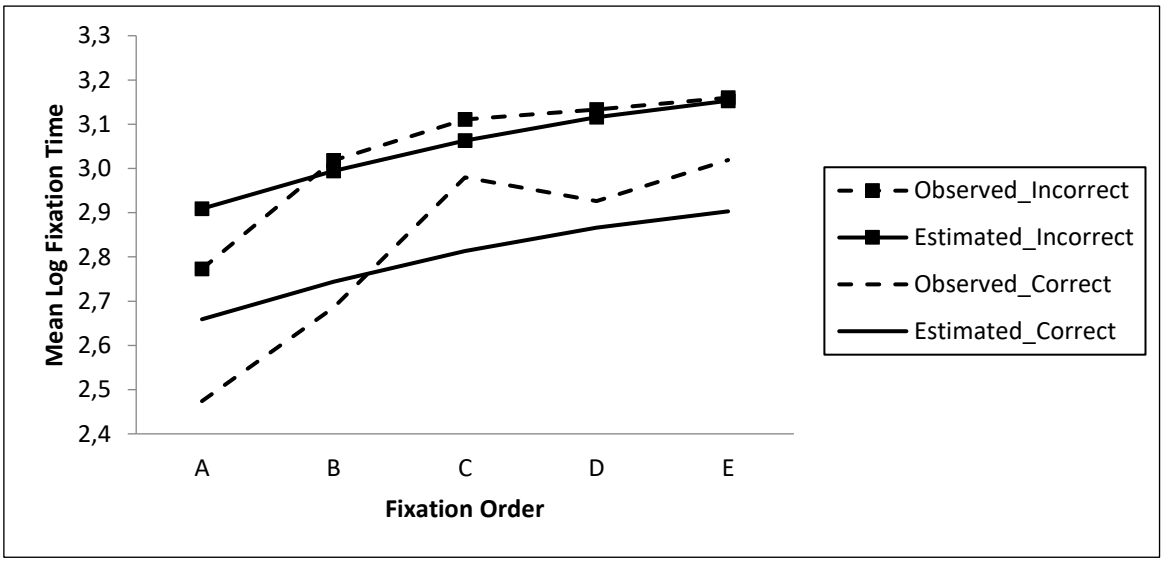


Figure 10 illustrates the group-specific random coefficients (intercepts and slopes) produced by the Quadratic Model with Correct/Incorrect Response Covariate (Model 3). Figure 10 demonstrates that the correct-response group and incorrect response group differed in two respects: in the observed scores of their initial status (intercepts) and their rates of change during the fixation orders (slopes). While there is a downward trend in mean log fixation times after the correct choice C for the correct-choice group, there is an upward trend in the mean log fixation times from choice A to choice E for the incorrect-choice group.

To sum up, subjects' reading behavior for choices was tested on two multiple-choice reading comprehension questions with a series of latent longitudinal models (Linear, quadratic and quadratic with a time-invariant covariate). For both item 1 and item 2, it was found that mean log fixation times increased quadraticly for subjects while reading over the choices A to E, meaning that subjects' reading speed was faster at first and slowed toward the final choice, choice E. The fact that the Quadratic model was reasonably acceptable for both items supports the finding in the previous sentence. Nonetheless, a better fit was observed for the quadratic model when the correct/incorrect response variable was added to the model as a time-invariant covariate. This finding reveals that the quadratic pattern estimated for the correct responders' was different than that of the incorrect responders, suggesting meaningful differences in both initial status and the change parameters for the two groups. These results are consistent with the expectation and imply that there were less of a searching behavior for subjects responding the item correctly after they reached to the correct coded choice.

Discussion and Conclusion

Eye-tracking is a technology that demonstrates great potential in educational assessment research because it provides time-series data that can be translated to item response process data for investigating the nature of the relationships between various cognitive processes and the performance of test-takers when expected to use these cognitive processes and has the potential to uncover the moment-to-moment processes of problem-solving behavior. Eye movements obtained from the eye-tracking devices are widely accepted to reflect cognitive processes for reading comprehension (Raney et al., 2014; Meziere et al., 2021); however, these cognitive processes cannot be directly inferred from eye-tracking data alone. In order to interpret eye-tracking data properly, theoretical and psychometrical models must always be the basis for designing experiments as well as for analyzing and interpreting eye-tracking based measurements.

Written from a psychometric perspective, this study illustrates that the first step of reaching accurate interpretations using eye-tracking enhanced data is to screen and translate time-series eye-movement data into item process data quantifying item surfaces or sections using binary-coded Areas of Interest variables. Within the context of this study, the Areas of Interest were multiple-choice item parts, each line in the question stem, and each choice. The results showed that the measures obtained from the subjects varied significantly in these areas when built as variables. For example, there was a considerable variation in how much time the subjects spent answering the items. Similarly, there was an apparent variation in how much time the subjects spent on each choice and each line, especially when the standard deviations were taken into account. This was taken as an indicator of the feasibility of the Areas of Interest utilized in this study.

In this study, special emphasis was given to the screening of eye-tracking recordings of subject-item encounters to spot recordings that were unreliable (eye movement recordings were not reliable for some subjects). The accuracy rate was approximately %85 and %92 for the two items studied in this study. This exemplifies that there could still be mismatches remaining, even after initial calibration tests (most, if not all, eye tracking devices have a calibration stage) were passed successfully. In order to overcome the problem of eye-gaze agreements being less than 100% accurate, our results suggest that, using a set of initial screening processes may help (involving analyses of graphs, scan paths, and heatmaps). Although ensuring reliability of the utilized eye movement data was of an interest in this study, our main goal was to use a novel approach, namely Latent Growth Modeling, to interpret the information that

might be available in item level eye movement data as it pertained to the particular choices presented. This was accomplished by binary coding time stamped fixations observed for each choice throughout examinee-item interaction time. Our findings reveal that through binary coding choice relevant time stamped eye movement data for choices A to E, it would be possible to map how fixation times (when ordered by choices A to E) changed over item-encounter time for each and all examinees. Albeit beyond the scope of this study, establishing a baseline trajectory for test items through the use of LGMs would be useful for flagging subjects responding an item correctly, yet, not showing any observed reading activity on any of the choices (possible cheating behavior). Another use could be for investigating item speededness by investigating if majority or a subgroup of examinees are running out of time before getting to the latter choices or if pseudo guessing has occurred.

To collect further evidence supporting validity arguments to be made about the accuracy of response item and test scores, an eye-tracking measure, total fixation duration, was calculated for the all the determined areas of interest (direction, lines and choices). It was clear that the test-takers paid their attention to these areas in a varying amount of time. In terms of the focus of this study, the fixation duration of the participants on each option were scrutinized, which shows that participicants who responded the item correctly spent more time on the correct answer/choice than the other choices (distractors). Also they fixated on the correct choice more than the participants who answered the item incorrectly, which is an indication of viewing inportant and relevant informaiton. Besides, participants gave more attention to some options (distractors) and less to others, which shows that some of the distractors were more related to construct to be measured while the others not. Of course, the fixation durations of partipicants on the options do not tell us the cognitive processes a participant uses while responding a multiple-choice quesiton; however, the distinction between these groups provides useful information on test validation.

The results from the latent growth models show that eye-tracking measurements obtained not only for the correct coded but also for the distracting choices may play an essential role in revealing the nature of within- and between-subject differences in reading behavior. Overall, the findings of this study show that the data obtained from the eye-tracking technology can be potentially useful in determining the patterns in reading behavior of subjects when responding to test items and testing if certain variables of interest, such as the response being correct or incorrect, explain some of the variances. Although not used in this study, model extensions can be easily formulated to include piecewise or cubic terms as well as additional time-invariant covariates such as gender, total score, etc., and time-varying covariates such as word count.

The scope of this study is limited in that it uses a small subject sample and only two multiple-choice item experiments. Another limitation is that this study focused on the fixation metrics while several other metrics are also provided by the eye-tracking technology, such as saccades. In addition, the caveat of this paper is that it uses a more of a psychometric perspective than a substantive one and that the future studies should focus more on the substantive issues. For future studies, researchers are recommended to use larger subject samples, and increased item experiments so that the placement order of the correct coded choices would not be limited. Albeit limited, the results of this study nicely illustrate that an effective integration of eye-tracking data into correct/incorrect response data may greatly enhance what we know about the item processing behaviors of subjects. The multi-stage data analysis approach was greatly useful for the findings from the initial screening and cross-sectional analyses were great pointers without which the longitudinal models could not have been easily estimated and interpreted. With larger sample sizes, multi-group latent growth curve models can be estimated to look into item and response reading patterns of subjects marking any of the five choices as the correct answer for a more detailed description of the subgroups being drawn to particular choices.

Declaration

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

Ethical Approval: The study was approved by the Gazi University Ethics Committee (Research code: 2019-003, dated 08.01.2019/ 01)

This paper presents some of the results obtained during the Doctoral Thesis process that were partially supported via the BAP Project No 04/2019-01 under the supervision of Prof. Dr. Nilüfer Kahraman.

Author Contribution: Ergün Cihat Çorbacı-Conceptualization, implementation, methodology, analysis, writing & editing. Nilüfer Kahraman-Conceptualization, methodology, analysis, writing & editing, visualization.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

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

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

Bentler, P. M. (1995). *EQS structural equations program manual* (vol. 6). Multivariate software. https://doi.org/10.4236/am.2014.510132

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136−162). Sage.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications*. Routledge.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

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

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In *Validation of score meaning for the next generation of assessments* (pp. 11-24). Routledge.

Mézière, D. C., Yu, L., Reichle, E., von der Malsburg, T., & McArthur, G. (2021). *Using eye-tracking measures to predict reading comprehension*. https://doi.org/10.31234/osf.io/v2rdp

Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 291–322). American Psychological Association. https://doi.org/10.1037/10409-010

Öğrenci Seçme ve Yerleştirme Merkezi (2018). Retrieved on February 25, from https://www.osym.gov.tr/TR,15313/2018-yds-sonbahar-donemi-temel-soru-kitapciklarinin-yayimlanmasi--10.html

Paulson, E. J., & Henry, J. (2002). Does the Degrees of Reading Power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent & Adult Literacy*, *46*(3), 234-244. https://link.gale.com/apps/doc/A94123361/LitRC?u=anon~924799ab&sid=bookmark-LitRC&xid=2cf242f9

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling* (No. 157). Sage.

Raney, G. E., Campbell, S. J., & Bovee, J. C. (2014). Using eye movements to evaluate the cognitive processes involved in text comprehension. *JoVE (Journal of Visualized Experiments)*, *83*, e50780. https://doi.org/10.3791/50780

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422. https://doi.org/10.1037/0033-2909.124.3.372

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464. https://doi.org/10.1214/aos/1176344136

Solheim, O. J., & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elemantary Education*, *4*(1), 153-168. https://www.iejee.com/index.php/IEJEE/article/view/218

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173-180. https://doi.org/10.1207/s15327906mbr2502_4

Tai, R. H., Loehr, J. F., & Brigham, F. J. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research & Method in Education*, *29*(2), 185-208. https://doi.org/10.1080/17437270600891614

Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers & Education*, *58*(1), 375-385. https://doi.org/10.1016/j.compedu.2011.07.012

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change. *Psychological Bulletin, 110*, 268-290. https://doi.org/10.1037/0033-2909.116.2.363

Yaneva, V., Clauser, B. E., Morales, A., & Paniagua, M. (2022). Assessing the validity of test scores using response process data from an eye-tracking study: A new approach. *Advances in Health Sciences Education, Online First*. https://doi.org/10.1007/s10459-022-10107-9

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 212-231

An Investigation of the Effect of Missing Data on Differential Item Functioning in Mixed Type Tests*

Leyla Burcu DİNÇSOY** Hülya KELECİOĞLU***

Abstract

In this research, the aim was to examine the effects of Markov Chain Monte Carlo (MCMC), multiple imputation (MI), and expectation maximization (EM), all methods of coping with missing data in mixed type tests including dichotomous and polytomous items, on the differential item functioning (DIF). The study was carried out on a complete data set consisting of the scores of 1160 students who took booklet number 9 in the science test in Trends in International Mathematics and Science Study (TIMSS) 2019 and answered it in full. The conditions to be examined for the effectiveness of the methods were missing data mechanism (MCAR and MAR), DIF level (A, B, and C), and missing data rate (10% and 20%). Data were assigned to the missing data sets created by deleting data at different rates under the missing completely at random (MCAR) and missing at random (MAR) mechanisms over the aforementioned data set. DIF analysis was performed on all the data sets obtained with the poly-SIBTEST method using the MCMC, MI, and EM methods. The results obtained from the complete data set were then compared with the result implications from other data sets for reference. The study showed that the EM and MCMC methods performed better for the C-level DIF than the A and B levels in terms of all conditions examined. MI was observed to be the most successful method in determining DIF in items showing DIF in 10% and 20% MCAR mechanisms. When compared with the complete data set, the three methods showed similar results in the 10% MAR mechanism while MCMC gave the closest results in the 20% MAR mechanism.

Keywords: mixed type test, missing data, differential item functioning, poly-SIBTEST

Introduction

The need to evaluate many individuals in a short period of time led to the development of achievement and aptitude tests in education (Özgüven, 2017). The assessments made as a result of these tests need to allow valid interpretations because they have individual, social, and political consequences. At this point, the problems of bias and missing data, which may lead to mistaken interpretations of the test results gain in importance.

It is essential that such achievement and aptitude tests do not contain bias and that they are fair for all those being tested. The technique known as DIF analysis is used in to statistically process bias (Zumbo, 1999). DIF refers to the differing performances of students with the same ability on an item. For an item that does not show DIF, it is expected that individuals with the same ability levels have the same probability of responding to the item correctly even if the individuals belong to different groups. However, if different item difficulties are observed in different groups with the same ability levels, the item exhibits DIF (Millsap & Everson, 1993). Camili (2006) examined DIF determination methods in two different groups: methods that equalize individuals by using observed scores and methods based on item response theory. Hambleton et al. (1993) divided DIF methods into three categories: those based on classical test theory, those based on item response theory, and those based on chi-square. Different researchers have classified DIF detection methods in different ways. Methods based on the classical test theory include Mantel-Haenszel (MH), logistical regression (LR), and the standardization method. Meanwhile, methods based on item response theory include Lord's chi-square, the likelihood ratio, and

* This study is a part of master's thesis conducted under the supervision of Prof. Dr. Hülya KELECİOĞLU and prepared by Leyla Burcu DİNÇSOY.

** Teacher, Ministry of National Education, Ankara-Turkey, leylaburcuadak@gmail.com, ORCID ID: 0000-0002-5633-3520
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Dinçsoy, L. B., & Kelecioğlu, H. (2022). An investigation of the effect of missing data on differential item functioning in mixed type tests. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 212-231. https://doi.org/10.21031/epod.1091085

Received: 21.03.2022
Accepted: 16.09.2022

Raju's area measurements (Camili, 2006; Zumbo, 1999). Many methods originally developed for items scored in two categories have later been expanded to include items scored in multiple categories. Some of these methods are MH, LR, and SIBTEST (Swaminathan & Rogers, 1990). This study used Poly-SIBTEST, the generalized version of the SIBTEST method that can be used for two- and multi-category data.

DIF may adversely affect the reliability of test scores as it may result in erroneous assessments for test takers. However, DIF is not the only factor that can lead to the increased validity of test scores. For instance, missing data is another such factor. It is also possible for both the DIF and the missing data to occur simultaneously (Garret, 2009). A review of the literature reveals studies whose primary aim was to determine the methods of coping with missing data that show better performance under different conditions (such as sample size, focus-reference group ratio, missing data rate, missing data mechanisms, polytomous-dichotomous items, mixed type tests, etc.), (Banks & Walker, 2006; Emenogu et al., 2010; Falenchuk & Herbert, 2009; Finch, 2011a; Garrett, 2009; Nichols et al., 2022; Sedivy et al., 2006; Tamcı, 2018) or the aim was to compare the performances of DIF detection methods in the presence of missing data (Finch, 2011b; Garrett, 2009; Robitzsch & Rupp, 2009; Rousseau et al., 2006; Sedivy et al., 2006). For example, Finch (2011a) examined the effectiveness of the DMF method in the presence of missing data in the uniform DMF analysis by calculating type 1 error and power ratios. He worked with items scored in two categories. He determined three different sample sizes and kept the focal reference group ratio constant. Under TRK, RK and ROK mechanisms, 5% and 10% of missing data were generated. It compared the effectiveness of the DMF, list-based deletion and zero assignment methods with MH, LR and SIBTEST methods by determining the DMF. He stated that the type 1 error rates for zero assignment are inflated in the RK mechanism, while the results in the TRK and ROK mechanisms are similar to the full dataset results. It is also stated that the type 1 errors and power ratios of the list-based deletion and RTA methods show similar results to the full dataset. Selvi and Alıcı (2018) examined the effect of missing data assignment methods on different DMF detection methods. The test consisting of eighty multiple-choice items was scored in two categories. In their study, BM and regression assignment were used as missing data handling methods, and MH, standardized method and likelihood ratio test were used as DMF detection methods. It was reported that the missing data assignment methods caused a difference in DMF items and this difference was significant in the MH method.

Deficiencies in the information collected for assessment cause a decrease in reliability and validity and increase the probability of inaccurate decisions (Turgut & Baykul, 2012). Missing data can occur for many different reasons. For instance, participants may deliberately choose not to respond to an item, overlook the item, forget to return to a skipped item, avoid answering the item, or not know the answer. Alternatively, the interviewer may skip the item, or it may not be suitable for the participant. The participant may even have to leave the study, or ultimately, errors might be made in the data entry stage (Allison, 2002). The assumption of almost all statistical methods is that all participants have complete information for the variables to be included in the analysis (Allison, 2002). For example, DIF detection methods such as Mantel-Haenszel (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), and simultaneous item bias test or SIBTEST (Shealy & Stout, 1993) are not designed for datasets with missing data. When the solution is the exclusion of participants with missing responses from the analysis, a large reduction in sample size may result, and detection power may be restricted if DIF is present (Banks, 2015). Instead, there are methods that researchers can choose to impute values to replace missing data. These methods have a significant impact on statistical results (Garret, 2009). Incorrect method selection can be a source of bias as it may result in masking the actual DIF or generating incorrect DIF in items that are not actually DIF (Banks, 2015). In the presence of missing data in the data set, the missing data mechanisms should be examined. The missing data mechanism is the mathematical relationship between the variables and the probability of the data being missing (Enders, 2010). Missing data mechanisms generally fall into three categories, classified by Little and Rubin (2020) as missing completely at random (MCAR), missing at random (MAR), and missing at not random (MNAR). This classification is the most widely accepted. In the case that a Y variable has missing data, in order to be able to state that the missing data in the Y variable are in the MCAR mechanism, the probability of missing data in the Y variable should be unrelated to the Y variable and

other variables. The MCAR mechanism is when the probability of missing data in the Y variable is irrelevant to the Y value when all other variables in the analysis are controlled (Allison, 2002). The ROK mechanism, on the other hand, states that the probability of missing data in the Y variable is related to the Y variable even after other variables in the dataset are controlled. The probability of missing data depends on the missing variable (Enders, 2010). Alpar (2021) summarized these mechanisms with an example as follows: In a study where weight was examined with the gender variable, it might be said that the data is in the MCAR mechanism if there is no reason among all the participants who did not state their weight, the MAR mechanism if the rate of women not answering their weight is higher, and the MNAR mechanism if those with more or less weight did not answer their weight.

In this way, missing data encountered in different mechanisms posed problems in analysis, which led researchers to search for solutions. This search for solutions started in the 1930s but became popularized with the work of Rubin (1976) (Toka, 2012). In general, methods of missing data imputation are grouped under two categories: deletion and simple imputation and probabilistic and offset data imputation. Probabilistic and translational methods are further divided into two groups as those based on the maximum likelihood approach and the MI approach (Demir, 2013). Examples of methods based on multiple data imputation approaches include listwise deletion, pairwise deletion, mean value imputation to deletion, and simple assignment-based methods; EM algorithm to methods based on maximum likelihood approach, direct maximum likelihood, and Bayesian data imputation methods; and MI, random imputation, and MCMC methods. MCMC, one of the methods used in this study, produces chains where each of the simulated values is lower than the previous value, unlike standard Monte Carlo methods, which generate a series of independent values through a simulation from the desired probability distribution. A Markov chain is a stochastic process with the property that any value in the value imputation sequence depends only on the previous value in its chain, thus being independent of all other prior states. The basic principle of MCMC is that when this Markov chain goes through a sufficient number of iterations, it will reach the desired posterior distribution (Gill, 2002). EM, on the other hand, is an iterative method that makes maximum probability estimations in two steps: expectation and maximization. This method begins first with the estimation of the mean vector and the covariance matrix. To estimate the missing data from the variables observed in the expectation step, a set of regression equations is set up using the mean vector and covariance matrix. By means of the established regression equations, a value is assigned to the missing data (Enders, 2010). Another method, MI, is designed to make multiple imputations instead of assigning a single value to the missing data and creates more than one complete data set (Van Buuren, 2012). It is performed in three stages: assigning data m times for each missing data, applying standard analyzes with m completed data sets, and combining the obtained m analysis results (Alpar, 2021).

In attempts to eliminate the problems and biases caused by missing data through data imputation, unconsciously imputed data does not eliminate the problem and may damage the reliability of the results. (Çüm et al., 2018; Little & Rubin, 1987). Therefore, it is important to determine the effect of value imputation methods instead of missing data in the presence of DIF in order to reduce the possible threats of DIF and missing data on validity of the test.

Purpose of the Research

A large number of studies can be found that deal with imputing data or DIF determination instead of missing data, but studies that deal with both are limited. Examining the literature and focusing on the effect of missing data on differential item functioning reveals that most studies have been conducted on dichotomous simulation data (e.g., Banks & Walker, 2006; Emenogu, 2010; Falenchuck & Herbert, 2009; Finch, 2011a, 2011b; Nichols et al., 2022; Robitzsch & Rupp, 2009; Rousseau et al., 2006). There are also some studies in which polytomous data are used (e.g., Garrett, 2009; Sedivy et al., 2006), but there are few studies using real data (e.g., Raousseau, 2004; Selvi & Alıcı, 2018; Tamcı, 2018). Because these studies generally use simulated data, tests in polytomous and dichotomous items are used separately. Therefore, this study examined the effects of missing data imputation methods on DIF under different conditions on the real data set in a mixed type test containing both polytomous and

dichotomous items. Since mixed type tests are frequently encountered in practice, new studies with mixed type tests are necessary.

This study is essential and will contribute greatly to the literature because it has been carried out with mixed type tests and real data, and it examines the effects of imputing values on real mixed test data containing missing data at two different rates and in two different missing data mechanisms using different methods on the differential item functioning in comparison with complete data sets. It is also important to see the effectiveness of different methods selected in the different conditions determined in the study.

The DIF analysis process is affected by missing data, as is the case with many analyses. If there are missing values in the dataset, appropriate methods should be selected accordingly, and there an effort must be made to prevent any problems that may be caused by the missing data. In this case, it is important to evaluate how much DIF results are affected by the method used and how similar the results it provides are to the real situation. If the missing data are not successfully compensated for, an item with DIF may appear as being without DIF because of the imputation method or an item without DIF may show DIF. Likewise, changes may occur in the DIF levels of items with DIF.

In order for the tests to give valid and reliable results, it is important to use the missing data imputation method that best compensates for these situations. As stated by Banks and Walker (2006), Finch (2011a, 2011b), and Garrett (2009), it is important for researchers to use one of these methods since when appropriate value imputation methods are used, results similar to full data sets are usually obtained.

In this regard, the research questions addressed in this study are:

1. How are the DIF results obtained by imputing data by the MCMC, MI, and EM methods to the data sets created by deleting 10% and 20% data in accordance with the MCAR mechanism from the full data set obtained from the TIMSS 2019 Science test?

2. How are the DIF results obtained by imputing data by MCMC, MI and EM method to the data sets created by deleting 10% and 20% data in accordance with the MAR mechanism from the full data set obtained from the TIMSS 2019 Science test?

3. How is the distribution of DIF results obtained by imputing data with MCMC, MI and EM methods to the data sets created by deleting 10% and 20% data in accordance with MCAR and MAR mechanisms from the full data set obtained from the TIMSS 2019 Science test differ according to the items showing and not showing DIF?

Method

The Model of the Research

This study was carried out with a correlational study model to examine the effect of distinctive methods of coping with missing data on DIF using reference results obtained from complete datasets in different circumstances. While the survey method describes the existing situation, correlational studies examine how the variables are related to each other (Karasar, 2011).

Participants

The data used in the study were obtained from the responses of students who participated in the TIMSS 2019 study conducted by the International Association for the Evaluation of Educational Achievement (IEA). The population of the study, therefore, consisted of approximately 250 thousand students who participated in the eighth grade TIMSS 2019 assessment. The sample was made up of students who took booklet number 9 in the eighth grade TIMSS 2019 evaluation and from the five native or non-native English-speaking countries where science score averages are close to each other, and thus there can be

no source of DIF. All the booklets were examined, and the booklet number 9, which contains the highest number of polytomous items, was selected.

The student answers with missing data were removed, and 1160 students were included in the analysis. The distribution of the number of students in the data set according to the native language variable was examined as a source of DIF, and the science score averages of the countries and their native languages are provided in Table 1.

Table 1

Distribution of Students Who Took Booklet Number 9 in the TIMSS 2019 Science Test by Countries, Science Averages, and Languages of the Countries

Countries	Number of Students	Science Average*	Language
England	177	517(4.8)	English
America	403	522(4.7)	English
Sweden	192	521(3.2)	Swedish
Turkey	194	515(3.7)	Turkish
Portugal	194	519(2.9)	Portuguese

*Standard errors are given in parentheses ().

Data Collection Tools

The research data consisted of the responses given by students from England, America, Sweden, Portugal, and Turkey to the seventeen items in booklet number 9, where the polytomous items of the TIMSS 2019 science test were the highest. Twelve of the 17 items used in the booklet were multiple-choice, and five were open-ended. Open-ended items were polytomous items scored as 0-1-2. In order to limit the research, the sample size and focus-reference group ratio were kept constant in the study.

Table 2

Examined Conditions

Conditions	Levels
DIF Level	A
	B
	C
Missing Data Rate	%10
	%20
Missing Data Mechanism	MCAR
	MAR
Imputation Methods	MCMC
	EM
	MI (5 imputation)

Data Analysis

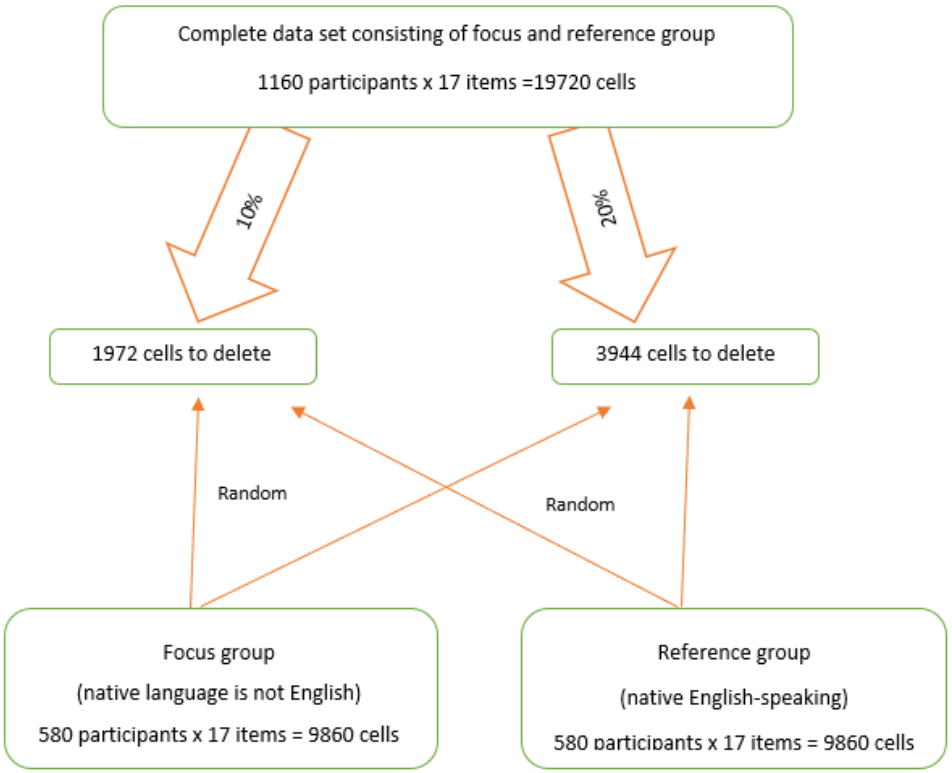
Creating Missing Datasets

Using the R program "missMethod" package from the complete dataset, four missing datasets were created in the MAR and MCAR mechanisms at rates of 10% and 20%. Missing data mechanisms constitute a bigger issue than the amount of missing data does. Though it is not an exact criterion, a missing rate of 5% or less with a random mechanism is negligible in large data sets (Tabachnick & Fidell, 1996). There are no strict rules regarding the negligibility of missing data with respect to their amounts. Similar studies in the literature show that these rates generally vary between 5% and 30%, and, in this study, the percentages were determined by taking these rates into account. The convenience of

the generated missing datasets for the missing data mechanism was examined using the IBM SPSS 24.0 program. The two datasets obtained in accordance with the MCAR mechanism were created by deleting an equal number of random data from each of the 17 items at rates of 10% and 20% from the complete dataset. The calculations related to the data deletion process in the MCAR mechanism are provided in Figure 1.

Figure 1

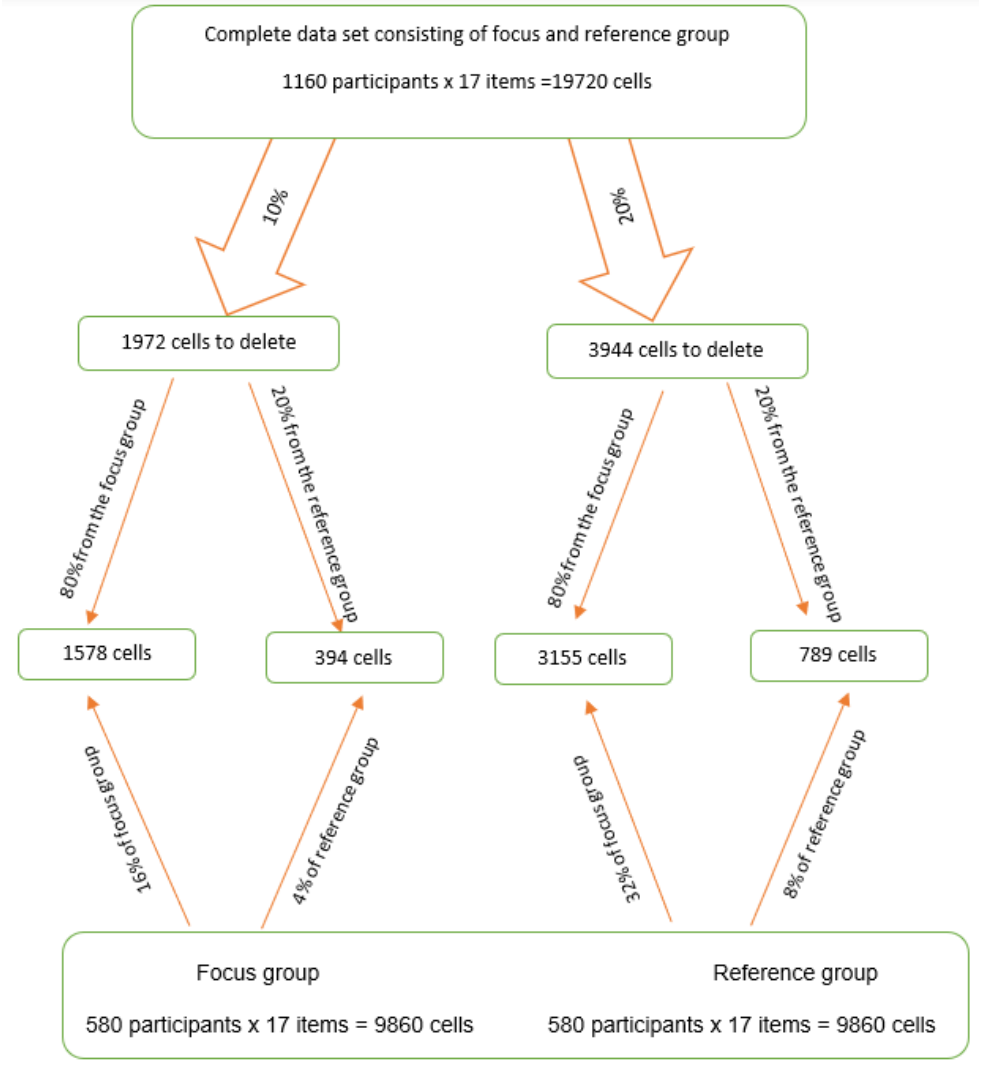
Deleting Missing Data Using the MCAR Mechanism



The process of generating missing data properly with the MAR mechanism was substantiated by deleting 80% of the data from the focus group and 20% of the data from the reference group randomly at rates of 10% and 20%. Calculations related to the data deletion in the MAR mechanism can be found in Figure 2.

Figure 2

Deleting Missing Data Using the MAR Mechanism



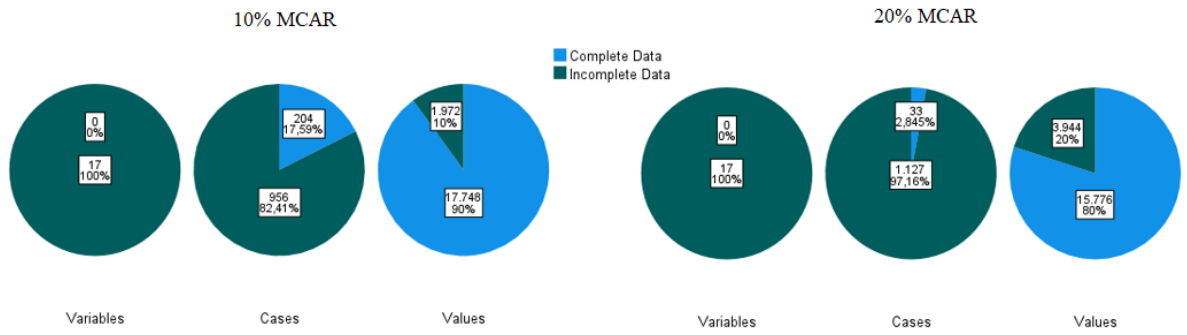
Examination of Missing Data Mechanisms

The examination of missing data mechanisms and imputing values instead of missing data were carried out on four datasets obtained at the previous stage. When the descriptive statistics of the datasets were examined before moving on to the missing data mechanism, no missing data for the native language variable was found in any datasets. In the two missing datasets created at 10%, 116 pieces of missing data were observed in each of the 17 items, a total of 1972 pieces of data. In the datasets created at the rate of 20%, 232 pieces of missing data were observed in each of the 17 items, and a total of 3944 pieces of data were examined.

The numbers and percentages of the missing data, which were determined by descriptive statistics in the variables, participants and values, are provided in the pie charts. Figure 3 shows the number of datasets with 10% and 20% missing data created by the MCAR mechanism in variables, participants, and values and their percentages.

Figure 3

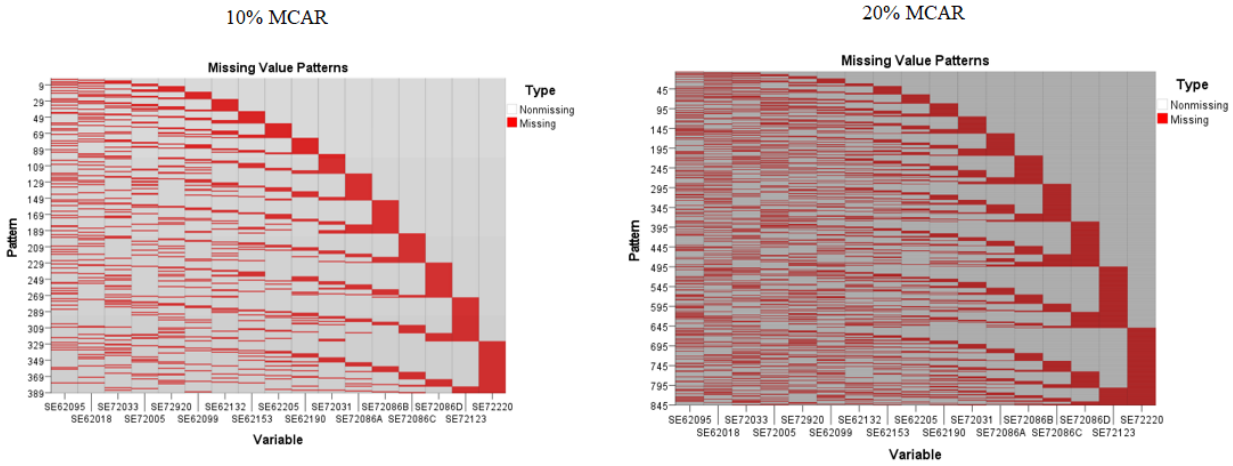
The ratio of Missing in the MCAR Dataset of 10% and 20% of the Data in Variables, Participants, and All Values



When Figure 3 is examined, it may be seen that there were missing data in all 17 variables (100%) in the variables graph in the 10% MCAR dataset, missing data were observed in 204 (17.59%) out of 1160 participants, and data were missing in 1972 (10%) of the 19720 cells in the data set. In the 20% MCAR data set, there were missing data in all 17 variables (100%) in the variables graph, missing data were observed in 33 (2.84%) out of 1160 participants, and data were missing in 3944 (20%) of the 19720 cells in the data set. Graphs showing missing data patterns are presented in Figure 4.

Figure 4

10% and 20% MCAR Dataset Missing Data Pattern Graphs



There are 389 and 845 patterns belonging to 17 variables and 1160 participants, respectively, at missing data rates of 10% and 20%. Since the gray cells representing the observed data are not clustered in the lower right part of the graph and the red cells representing the missing data are not clustered in the upper left part of the graph, it can be stated that the missing are not in any specific order and that there is a non-monotonous pattern.

Whether the missing in the missing datasets have the MCAR mechanism was also examined by the Little & Rubin (2020) MCAR test. Since the p values for both datasets were not statistically significant (p=0.099 for a 10% missing dataset, p=0.656 for a 20% missing dataset), this served as evidence that the data were completely random. Although there are statistical tests regarding the compatibility of the missing data for the MCAR mechanism, this is not the case for the MAR mechanism. In this case, characterizing the data as MAR is only an assumption (Schafer & Graham, 2002).

Imputation Methods Instead of Missing Data

The missing data were imputed to the created datasets using the MCMC, EM, and MI methods. The MCMC imputation was performed with Lisrel 8.80, and the imputations were performed with the EM and MI methods with the IBM SPSS 24.0 program. Eight datasets were generated by imputing four datasets with the rates of 10% and 20% missing data in the MCAR and MAR mechanisms using the EM and MCMC methods. The number of imputations to be performed by the MI method is determined in correlation to the efficiency table of the number of imputations that can be made for distinctive missing data rates (Schafer & Olsen, 1998).

Table 3

Relative efficiency of the number of assignments at different Missing Data Rates

Number of	Missing Data Ratio				
Assignments	%10	%30	%50	%70	%90
3	.97	.91	.86	.81	.77
5	.98	.94	.91	.88	.85
10	.99	.97	.95	.93	.92
20	1.00	.99	.98	.97	.96

When Table 3 is examined, it is seen that when the missing data is at low rates, the effect of the number of assignments is low, while the number of assignments becomes more important as the rate increases. For example, the difference between 3 assignments and 10 assignments at a rate of 10% can be interpreted as there is no need to make many assignments.

Twenty datasets were generated by imputing five data to each of the four datasets using the MI method. A total of twenty-eight datasets were created, which were imputed by three methods. Of these datasets, 20 datasets imputed by the MI were combined with DIF analysis.

Differential Item Functioning Analysis

In the final stage, DIF analysis was performed regarding the complete dataset and the native language variable. The focus group consisted of participants whose native language was not English, and the reference group consisted of native English-speaking participants. The descriptive statistics of the complete dataset are given in Table 4.

Table 4

Descriptive Statistics of the Complete Dataset

Statistics			
N_O, N_R	580	Mean	10.90
N_T	1160	Mode	12
$K_İ$	12	Median	11
$K_Ç$	5	Standard Deviation	4.27
K_T	17	Skewness	-0.021
Minimum Score	0	Kurtosis	-0.554
Maximum Score	22	Cronbach's Alpha	0.758

Notes: Focus group sample size, N_R: Reference group sample size, N_T: Total sample size, $K_İ$: Number of items dichotomous, $K_Ç$: Number of items polytomous, K_T: Total number of items

Examining the descriptive statistics showed that there were 1160 participants in total, including 580 from countries whose native language was not English in the focus group and 580 from countries whose native language was English in the reference group. There were five polytomous items (SE62095, SE62018, SE72033, SE72005, SE72920) and 12 dichotomous items (SE62099, SE62132, SE62153, SE62205, SE62190, SE72031, SE72086A, SE72086B, SE72086C, SE72086D, SE72123, SE72220) in the dataset. The mean of the booklet was 10.9, the mod was 12, the median was 11, and since they are close to each other, it can be stated that the distribution is quite normal. The skewness and kurtosis coefficients are in the range of +1, -1, which can be interpreted as not deviating excessively from the normal.

DIF analysis was performed on 28 datasets obtained by imputing values instead of missing data. The Poly-SIBTEST technique was used in the analysis. SIBTEST is a DIF determination method developed by Shealy and Stout (1993) in data scored dichotomously. Poly-SIBTEST is the general version of the SIBTEST method that can be used for dichotomous and polytomous data (Fang, 1999). Using the Poly-SIBTEST technique, DIF can be used to analyze both the item packages and the items in the data set one by one (Camilli, 2006). Since the study did not compare DIF methods, it was considered appropriate to choose a single DIF determination method. Banks and Walker (2006) used only SIBTEST in their study, obtaining good results. Sedivy et al. similarly determined DIF using the pol-SIBTEST method, stating that this was a suitable method for tests containing polytomous items. Test items are divided into two subtests, matching items and suspect items. Matching items are used as an internal matching criterion to check for group differences in target ability that is intended to be measured in DIF detection (Bolt, 2000). The analysis is carried out by dividing the test into two so that the items to be analyzed for DIF are taken into one group, and the remaining items are taken into the second group. In order to compare the performances on the items studied in the DIF, matching is done over the actual scores estimated by the total scores on the items in the second group (Gierl, 2005).

The estimate of the Poly-SIBTEST DMF index is given as

$$\hat{\beta} = \sum_{k=0}^{n_m} p_k \, (\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

with k being the number of score levels in the matching items, n_m being the maximum score level in the current matching items, pk being the proportion of individuals with k matching items, Y_{Rk} being the average item score over the working group for the reference group at the current matching items, and Y_{Fk} being is the average item score for the focus group over the working group at the current matching items level.

The poly-SIBTEST statistic associated with the null hypothesis is the same as the test statistic used with the SIBTEST procedure.

$$poly - SIBTEST = \hat{\beta} \, / \hat{\sigma}(\hat{\beta})$$

$\hat{\sigma}(\hat{\beta})$ indicates that the Poly-SIBTEST DMF index is the estimated standard error.

The criteria proposed by Roussos & Stout, (1996) in interpreting the DIF effect size obtained with SIBTEST and Poly-SIBTEST are shown in Table 5. When this value is negative, the item shows DIF in favor of the focus group, and when it is positive, it shows DIF in favor of the reference group.

Table 5

β Values Interpretation Measures

DIF Level	β Value
A Level (can be ignored)	IβI <0,059
B Level (Medium Level)	0.059 ≤IβI ≤ 0,088
C Level (High Level)	IβI ≥ 0,088

DIF analyses were performed for each item one by one on 28 data sets obtained by complete data set and imputation. The reason the analysis was performed for each item separately is that the total score of the individuals in the reference and focus groups was formed by the matching items and that suspect items were not included in the matching variable. While the single item in the suspect items was analyzed, the other 16 items included in the matching items determined the total scores.
Since the number of imputations was high in the MI method, DIF analysis is given in stages.

1) With the MI method, five different imputations were made for each dataset, and 20 datasets were obtained.
2) DIF analysis was performed for the 20 datasets obtained.
3) The DIF analysis averages of the five imputations made for the missing dataset were combined.
4) The third process was repeated for all four datasets.

After DIF analysis of the data sets was obtained by imputing values instead of missing data, the results were compared with the results of the full data set DIF analysis and examined to see if there were differences in whether the items showed DIF or not and the changes in DIF levels.

Results

Findings Regarding the Complete Dataset

The poly-SIBTEST results obtained from the complete dataset to be referenced in the comparisons are presented in Table 6.

Table 6

poly-SIBTEST Analysis Findings of the Complete Dataset According to the Language Variable

Item No	β̂	p	DIF Level	Advantageous Groups	Item No	β̂	p	DIF Level	Advantageous Groups
1	-0.330	0.000*	C	Focus	10	0.169	0.000*	C	Reference
2	0.147	0.000*	C	Reference	11	-0.155	0.000*	C	Focus
3	0.212	0.000*	C	Reference	12	0.038	0.046*	A	Reference
4	0.123	0.002*	C	Reference	13	0.022	0.261		
5	-0.109	0.003*	C	Focus	14	0.008	0.671		
6	0.006	0.812			15	0.028	0.097		
7	0.042	0.133			16	0.130	0.000*	C	Reference
8	0.060	0.036*	B	Reference	17	-0.029	0.255		
9	-0.242	0.000*	C	Focus					

The complete dataset determined that one item showed DIF at A level, one item at the B level, and nine items at the C level. While items 1, 5, 9, and 11 were in favor of the focus group, the non-native English speakers, the other DIF items showed DIF in favor of the reference group, that is, native English speakers.

Findings Regarding Datasets Deleted by the Rates of 10% and 20% According to the MCAR Mechanism

According to the MCAR mechanism, the missing data obtained by deleting 10% and 20% of the data were imputed using the MCMC, EM, and MI methods. The DIF analysis of these data was performed with poly-SIBTEST, and the results are presented in Table 7.

Table 7

poly-SIBTEST Analysis Findings According to Language Variable of Datasets Generated by MCMC, EM and MI Imputation with MCAR Mechanism

Imputation Methods	MCMC				EM				MI			
Missing Data Rate	10%		20%		10%		20%		10%		20%	
Item No	β̂	p	β̂	p	β̂	p	β̂	p	β̂	p	β̂	p
1	-0.298	0.000*	-0.261	0.000*	-.291	0.000*	-0.264	0.000*	-0.285	0.000*	-0.254	0.000*
2	0.134	0.000*	0.133	0.000*	0.127	0.000*	0.122	0.000*	0.134	0.000*	0.133	0.000*
3	0.216	0.000*	0.140	0.001*	0.222	0.000*	0.150	0.000*	0.213	0.000*	0.144	0.002*
4	0.121	0.002*	0.100	0.006*	0.127	0.001*	0.097	0.006*	0.132	0.001*	0.093	0.113
5	-0.120	0.001*	-0.096	0.004*	-.128	0.000*	-0.102	0.002*	-0.130	0.002*	-0.104	0.004*
6	-0.017	0.532	0.006	0.830	-.024	0.370	-0.017	0.519	-0.010	0.906	-0.014	0.831
7	0.052	0.056	-0.005	0.842	0.054	0.046	0.013	0.614	0.042	0.489	0.060	0.246
8	0.058	0.039*	0.053	0.052	0.055	0.048*	0.046	0.091	0.048	0.034*	0.045	0.007*
9	-0.227	0.000*	-0.201	0.000*	-.223	0.000*	-0.195	0.000*	-0.223	0.000*	-0.191	0.000*
10	0.157	0.000*	0.127	0.000*	0.173	0.000*	0.144	0.000*	0.163	0.000*	0.126	0.000*
11	-0.137	0.000*	-0.133	0.000*	-.147	0.000*	-0.117	0.000*	-0.148	0.000*	-0.113	0.000*
12	0.027	0.145	0.021	0.246	0.025	0.183	0.026	0.155	0.031	0.047*	0.022	0.060*
13	0.024	0.202	0.015	0.443	0.022	0.245	0.014	0.454	0.020	0.247	0.016	0.306
14	0.017	0.380	0.000	0.988	0.009	0.614	0.000	0.988	0.014	0.376	-0.007	0.919
15	0.028	0.110	0.024	0.153	0.024	0.164	0.026	0.108	0.030	0.036*	0.020	0.271
16	0.122	0.000*	0.119	0.000*	0.111	0.000*	0.139	0.000*	0.116	0.000*	0.120	0.000*
17	-0.045	0.071	0.005	0.845	-.038	0.132	0.007	0.779	-0.046	0.068	-0.014	0.409

According to the poly-SIBTEST results of the data set generated by imputing MCMC to the 10% MCAR mechanism data set, it was observed that item 12 of the items showing DIF in the complete data set did not show DIF, and item 8 was a different DIF level. When compared with the results of the complete data set, items 8 and 12 of the items with DIF did not show DIF. In both of the data sets with missing data in both ratios and imputed with the MCMC method, none of the items without DIF in the complete data set showed DIF. Of the items with DIF in the complete data set, 91% and 82% showed DIF at 10% and 20%, respectively.

According to the poly-SIBTEST results of the data set generated with the EM imputation, the data set with the 10% MCAR mechanism showed that the 12th item among the items with DIF in the full data set did not show DIF, while the seventh item without DIF showed DIF, unlike the results of the full data set. In addition, the DIF level of item 8 showed a difference. The poly-SIBTEST results of the data set created by imputing EM to the 20% MCAR mechanism data set determined that nine items showed DIF at the C level, and unlike the results of the complete data set, the eighth and 12th of the items with DIF did not show DIF. In the complete data set of data sets, which contained 10% and 20% missing data and was imputed with EM method, 83% and 100% of the items without DIF did not show DIF. 91% and 82% of the items with DIF in the complete data set showed DIF, respectively, in the rates of 10% and 20%. According to the poly-SIBTEST results of the data set created with the MI imputation, it was determined that the 15th item, which did not show DIF in the full data set, showed DIF at the A level, and the eighth item had a different DIF level, unlike the full data set in the 10% MCAR mechanism data

set. In the data set generated by imputing MI to the rate of 20% MCAR mechanism data set, unlike the results of the complete data set, the fourth item with DIF did not show DIF, the 15th item without DIF showed DIF at the A level, and the level of DIF for the eighth item showed a difference. In both data sets with 10% and 20% missing data and imputed by the MI method, 83% of the items without DIF in the complete data set did not show DIF. Of the items with DIF in the complete data set, 100% and 91% showed DIF, respectively, at the rates of 10% and 20%.

Findings Regarding Datasets Deleted by the Rate of 10% and 20%

According to the MAR mechanism, data were imputed to missing data obtained by deleting 10% and 20% of them using the MCMC, EM, and MI methods. The DIF analysis of these data were performed with poly-SIBTEST, and the results are shown in Table 8.

Table 8

poly-SIBTEST Analysis Findings According to Language Variable of Datasets Generated by MCMC, EM, and MI Imputation with MAR Mechanism

Imputation Methods	MCMC				EM				MI			
Missing Data Rate	10%		20%		10%		20%		10%		20%	
Item	β	p	β	p	β	p	β	p	β	p	β	p
1	-0.298	0.000*	-0.291	0.000*	-0.288	0.000*	-0.278	0.000*	-0.304	0.000*	-0.304	0.000*
2	0.154	0.000*	0.136	0.000*	0.148	0.000*	0.159	0.000*	0.135	0.001*	0.110	0.004*
3	0.188	0.000*	0.177	0.000*	0.193	0.000*	0.177	0.000*	0.171	0.000*	0.191	0.000*
4	0.114	0.003*	0.118	0.001*	0.124	0.002*	0.131	0.000*	0.127	0.005*	0.085	0.028*
5	-0.093	0.008*	-0.120	0.000*	-0.107	0.002*	-0.148	0.000*	-0.088	0.015*	-0.104	0.010*
6	0.008	0.768	-0.018	0.501	0.002	0.944	0.018	0.499	0.013	0.871	0.024	0.607
7	0.012	0.668	0.046	0.089	0.017	0.531	0.017	0.519	0.027	0.189	0.053	0.030*
8	0.041	0.137	0.056	0.038*	0.036	0.185	0.056	0.035*	0.046	0.174	0.067	0.035*
9	-0.200	0.000*	-0.172	0.000*	-0.199	0.000*	-0.175	0.000*	-0.206	0.000*	-0.201	0.000*
10	0.151	0.000*	0.104	0.000*	0.148	0.000*	0.106	0.000*	0.150	0.000*	0.141	0.000*
11	-0.142	0.000*	-0.106	0.000*	-0.121	0.000*	-0.111	0.000*	-0.133	0.000*	-0.146	0.000*
12	0.028	0.145	0.001	0.941	0.017	0.390	-0.004	0.841	0.028	0.131	0.025	0.229
13	0.004	0.838	-0.008	0.687	-0.004	0.825	-0.013	0.484	0.011	0.326	0.011	0.341
14	-0.002	0.896	-0.014	0.460	-0.009	0.606	-0.018	0.333	0.006	0.693	0.008	0.702
15	0.023	0.180	0.012	0.497	0.014	0.412	0.002	0.887	0.029	0.112	0.010	0.819
16	0.090	0.001*	0.092	0.001*	0.095	0.000*	0.099	0.000*	0.101	0.000*	0.107	0.000*
17	-0.006	0.814	0.043	0.069	-0.005	0.851	0.072	0.002*	-0.033	0.092	-0.013	0.894

According to the poly-SIBTEST results of the data obtained by imputing MCMC to the data with 10% MAR, it was observed that the eighth and 12th items with DIF in the complete data set did not show DIF. When the results of the data set generated by imputing MCMC to the data set with a 20% MAR mechanism in the complete data set were examined, it was seen that the 12th item with DIF did not show DIF and the eighth item had a distinctive DIF level. In both of the data sets with 10% and 20% missing data and imputed with the MCMC method, none of the items without DIF in the complete data set showed DIF. Of the items with DIF in the complete data set, 82% and 91% of them showed DIF, respectively, at the rates of 10% and 20%.

According to the results of the poly-SIBTEST of the data set generated by the EM imputation, compared with the results of the complete data set, it was seen that the eighth and 12th items with DIF did not show DIF. It was determined that the 12th item with DIF did not show DIF, and the DIF levels of the eighth and 17th items showed a difference in the data set created by imputing EM to the 20% MAR

mechanism data set, unlike the results of the complete data set. In both data sets with 10% and 20% missing data and data sets that imputed with the EM method in order, all and 83% of the items without DIF in the complete data set did not show DIF. In both data sets, 82% of the items with DIF in the complete data set showed DIF.

According to the poly-SIBTEST results of the data set generated with the imputation of MI, in the 10% MAR mechanism data set, unlike the complete data set, the eighth and 12th items with DIF did not show DIF. In the data set generated by imputing MI to the data set with a 20% MAR mechanism, it could be seen that the 12th item with DIF did not show DIF, the seventh item without DIF showed DIF, and the fourth item had different DIF levels. While 83% of the items without DIF in the complete data set did not show DIF in both of the data sets consisting of 10% and 20% missing data and imputed with the MI method, the items with DIF in the complete data set were at the rates of 10% and 20%, respectively, and 100% and 91% of them showed DIF again. In the complete data set of the data sets with 10% and 20% missing data and imputed with the MI method in order; all and 83% of the items without DIF did not show DIF. 82% and 91% of the items with DIF in the complete data set showed DIF.

Findings Regarding the Distribution of Items Displaying and Not Displaying DIF

The distributions of the 12 items displaying DIF in the complete dataset as a result of imputing values using the EM, MCMC, and MI methods are presented in Table 9.

Table 9

Distributions of the Items Displaying DIF in the Complete Dataset According to the Missing Data Mechanisms and Missing Data Imputation Methods in the Missing Data

Missing Data Mechanism		MCAR						MAR					
Missing Data Imputation Method		EM		MCMC		MI		EM		MCMC		MI	
Missing Data Rate	DIF Level	f	%	f	%	f	%	f	%	f	%	f	%
10%	A	0	0	0	0	1	100	0	0	0	0	0	0
	B	1	100	1	100	1	100	0	0	0	0	0	0
	C	9	100	9	100	9	100	9	100	9	100	9	100
20%	A	0	0	0	0	1	100	0	0	0	0	0	0
	B	0	0	0	0	1	100	1	100	1	100	1	100
	C	9	100	9	100	8	89	9	100	9	100	9	100

A and B levels of DIF could only be determined by imputation with the MI method at a missing data rate of 20% under the MCAR mechanism. While the B level can be determined in all methods at the rate of 10% missing data under the MCAR mechanism, the A level could only be determined by the MI method. While under the MAR mechanism, items displaying DIF at levels A and B at a rate of 10% missing data could not be determined in all three methods, and only the A level could not be determined at a rate of 20%. In general, as the rate of missing data increased, the inability to correctly identify items with DIF increased. Similarly, Tamcı (2018) stated that the status of DIF items as a result of imputation with MI and EM methods shows good results in some circumstances while the status of items with DIF as a result of imputation was badly affected in an increase in the missing data rate. The items displaying DIF at the B and C levels had similar outcomes to the complete dataset in all methods at a missing data rate of 20% under the MAR mechanism and a missing data rate of 10% under the MCAR mechanism.

As for the C level DIF, the same results were obtained with the complete dataset in all other conditions except for the imputation by the MI method at a missing data rate of 20% under the MCAR mechanism. Under the MAR mechanism, all C-level DIF items were identified by all methods. When we look at the situation of whether the items displaying DIF still displayed DIF as a results of value imputation, it was found that 10% of the data were missing under the MCAR mechanism, and DIF was identified in all of

the items displaying DIF in the complete dataset at all DIF levels when the imputation was made with the MI method.

The distributions of the six items that do not display DIF in the complete dataset as a result of value imputation are given in Table 10.

Table 10

Distributions of the Items Not Displaying DIF in the Complete Dataset According to the Missing Data Mechanisms and Missing Data Imputation Methods in the Missing Data

Missing Data Mechanism	MCAR						MAR					
Missing Data Imputation Method	EM		MCMC		MI		EM		MCMC		MI	
Missing Data Rate	f	%	f	%	f	%	f	%	f	%	f	%
10%	5	83,3	6	100	5	83,3	6	100	6	100	6	100
20%	6	100	6	100	5	83,3	5	83,3	6	100	5	83,3

In all circumstances, the MCMC method correctly identified all items in the complete dataset that were not displaying DIF. DIF was observed in one item in which no DIF was observed in the complete dataset when imputing it using the EM and MI methods at a missing data rate of 10% under the MCAR mechanism and at a missing data rate of 20% under the MAR mechanism.
When examining whether items without DIF display DIF, the MCMC method was usually found to be preferable to the other methods in all conditions. Garrett (2009) stated that under the MCAR mechanism, the fact that items without DIF did not show DIF as a result of imputation with the MI method was better than the other methods used in their study.

Discussion and Conclusion

This study examined how DIF results differentiate according to the DIF level, missing data rate, and missing data mechanism when data imputation is performed using the MCMC, EM, and MI methods and considering the native language variable. The findings obtained from this study were presented by examining how the results obtained from the complete dataset and the MCMC, EM, and MI methods change in each missing data mechanism and missing data rate.

According to the missing data rate condition discussed in the research, it was seen that when the MCMC method was used, the increase in the missing data rate under the MCAR mechanism decreased in the correct identification of DIF, and the increase in the missing data rate under the MAR mechanism showed improvement in detecting DIF correctly. If there is an increase in the rate of missing data under the MAR mechanism, the use of the MCMC method was found to be more suitable. Finch (2011b) stated in her study that assigned stochastic regression imputation, logistic regression, and MI with zero imputation; for MI, the error decreased as the percentage of missing data increased while there was no decrease or increase in other methods. Nichols et al. (2022) stated that when the missing data rate is larger than 10% much larger magnitudes of DIF estimation error were observed.

Considering DIF levels, the MCMC and EM methods had good results in items with DIF at the C level under both the MCAR and the MAR mechanisms. While the MI method performed well with a 10% missing data rate, incorrect identifications were observed at the C level when the missing data rate increased. Based on these results, while all three methods can be preferred at a 10% missing data rate at a high DIF level (C), the EM and MCMC methods were found to be preferable to the MI method with a 20% missing data rate. The EM and MCMC methods with a missing rate of 20% under the MCAR mechanism and three methods with a missing rate of 10% under the MAR mechanism had the same results. In the MCMC and EM methods, it was observed that items with A and B levels DIF could not

be identified in the 20% MCAR and 10% MAR mechanisms, and while the A level DIF could not be identified in the 10% MCAR and 20% MAR mechanisms, it identified the B level as DIF at a lower level. While the MI method was some successful in identifying the A level in all missing data rates under the MCAR mechanism, it was determined the B level was identified at a low level, like other methods. The 10% MAR mechanism could determine DIF only at the C level and was unable to determine DIF at other levels. The finding was that the MI method is preferable to other methods for identifying A-level DIF. A level DIF was seen in items without DIF with the MI method in the MCAR mechanisms and 20% MAR mechanism.

With the increase in the rate of missing data in the MCAR mechanism, incorrect determinations were observed in the determination of substances with DIF. As the missing data rate increased, the EM and MCMC methods were found to be preferable over the MI method. Items with DIF were best identified under the MCAR mechanism when imputing them with the MI method at a missing data rate of 10%. While all items with DIF were determined by this method, the level of only one item with DIF was determined distinctively. Similarly, Finch (2011a) mentioned that the results obtained when he imputed using the MI method under the MCAR mechanism were compatible with the complete data, and, in his other study, Finch (2011b) stated that the results obtained when he imputed using the MI method again, except for the 10% missing data condition, were also compatible with the complete data. In the other study, Finch stated that when the type I error rates in the MAR mechanism were examined, the error rates for MI were lower than for the other two methods (zero imputation and stochastic regression imputation). On the other hand, Garrett (2009), in his study investigating the effects of MI and mean value imputing methods and the MH and ordinal LR methods, which are DIF determination methods suggested the use of MI, one of the methods of coping with missing data, when both DIF determination methods are used.

When the situations of whether items without DIF displayed DIF were examined, DIF was detected in items without DIF when using MI method at the rate of 10% and 20% missing data and EM methods at the rate of 10% missing data under the MCAR mechanism. Under the MAR mechanism, DIF was identified in items without DIF when EM and MI methods were used at a missing data rate of 20%. The MCMC method showed good results by identifying all items without DIF at both missing data rates under the MCAR mechanism, the EM method at a missing data rate of 20%, the MCMC method at both missing data rates under the MAR mechanism, and the EM and MI methods at a missing data rate of 10% without DIF. These indicators support the results of the study conducted by Tamcı (2019), which indicated that the MI method works better on items that do not display DIF than the EM method, and all items that did not display DIF for all other conditions at the missing data rate of 30% came out without DIF in the MI method.

In light of these results, it cannot be said that only one method is good. Different results were obtained for various conditions. For example, Finch (2011b) stated that study results suggested that the relationships between the different factors manipulated were complex with no one method emerging as fixed in all cases; however, listwise deletion consistently produced results similar to those obtained with the complete data set across simulated conditions. When the missing data rate exceeds the 10% threshold, Nichols et al. (2022) recommended MICE in their study due to missing data when testing for DIF. However, they stated that the methods were unable to completely eliminate the observed error due to missing data. Therefore, whichever method is used by the researchers, they should interpret the results carefully.

Recommendations Based on Research Results

Since the results of the MCMC and EM methods are found to be more similar to the complete dataset in cases where there are items displaying C-level DIF, these methods are recommended for imputing missing data in DIF studies. The MI method can be preferred when there are items with DIF at A and B levels.

In cases where the rate of missing data is high in the correct detection of items without DIF, it is recommended that the MCMC and EM methods be used under the MCAR mechanism and the MCMC method under the MAR mechanism.

For low missing data rates, it is recommended that the MCMC method be used under the MCAR mechanism and all three methods under the MAR mechanism.

In the identification of items with DIF, it is recommended that the MI method be used in cases where the missing data rate is high under MCAR and MAR mechanisms, and all three methods should be used with low missing data rates.

Recommendations Based on Subsequent Research

Based on the results of the research, in cases where there are items showing C-level DIF, the results of MCMC and EM methods are found to be more similar to the complete data set, so it is recommended that these methods be chosen for imputing missing data in DIF studies. When there are items with DIF at A and B levels, the MI method can be preferred.

In cases where the rate of missing data is high in the correct determination items without DIF, it is recommended that MCMC and EM methods be used under the MCAR mechanism and the MCMC method under the MAR mechanism.

For low missing data rates, it may be recommended that the MCMC method be used under the MCAR mechanism and all three methods under the MAR mechanism. In the detection of items with DIF, it can be recommended that the MI method be used in cases where the missing data rate is high under MCAR and MAR mechanisms, and all three methods should be used with low missing data rates.

In this study, a single DIF method was used. Despite DIF determination methods having similar results in general, as Gök et al. (2010) stated, there is no complete harmony between the methods, and it is recommended that different DIF methods be used, such as LR, the IRT probability rate, which has used different matching criteria, algorithms, and breakpoints.

In this study, the sample size was fixed. The mechanisms for missing data and methods of dealing with missing data at different sample sizes could be studied. Of the methods for dealing with missing data, three imputation methods were used from the class of probabilistic and translational data imputation. The number of methods can be increased, and comparisons can be made by using methods based on the deletion and simple imputation. DIF analysis was conducted using a CTT-based method. DIF analysis can also be conducted with IRT-based methods and techniques. In addition, the study can be expanded by increasing conditions for the test length, uniform and non-uniform DIF, focus-reference group rates, missing data rate, and methods for dealing with missing data.

Declarations

Author(s) contribution: Leyla Burcu DİNÇSOY-Investigation, methodology, visualization, software, formal analysis, and writing-original draft. Hülya KELECİOĞLU-Developing process, methodology, resources, supervision and validation.

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

Ethical Approval: This research study complies with research publishing ethics. Secondary data were used in this study. Therefore, ethical approval is not required.

References

Allison, P. D. (2002). *Missing data*. Sage.
Alpar, R. (2021). *Çok değişkenli istatistiksel yöntemler.* Detay.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, *20*(12), 12. https://doi.org/10.7275/FPG0-5079

Banks, K., & Walker, C. M. (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.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*(4), 307-327. https://doi.org/10.1111/j.1745-3984.2000.tb01089.x

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). American Council on Education & Praeger Publishers.

Çüm, S., Demir, E. K., Gelbal, S., & Kışla, T. (2018). A comparison of advanced methods used for missing data imputation under different conditions. *Mehmet Akif Ersoy University Journal of Education Faculty, 45*, 230-249. https://doi.org/10.21764/maeuefd.332605

Demir, E. (2013). Item and test parameters estimations for multiple choice tests in the presence of missing data: The case of SBS. *Journal of Educational Sciences Research, 3*(2), 47-68. http://dx.doi.org/10.12973/jesr.2013.324a

Emenogu, B. C., Falenchuck, 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.

Falenchuk, O., & Herbert, M. (2009). Investigation of differential non-response as a factor affecting the results of Mantel-Haenszel DIF detection. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Fang, T. (1999). *Detecting DIF in polytomous item responses* [Doctoral dissertation, University of Ottawa]. https://ruor.uottawa.ca/handle/10393/8495?locale=fr

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. W. (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

Garrett, P. (2009). *A Monte Carlo study investigating missing data, differential item functioning and effect size* [Doctoral thesis, Georgia State University]. https://doi.org/10.57709/1060078

Gierl, M. J. (2005). Using dimensionality-based DIF analysis to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*, 3-14. https://doi.org/10.1111/j.1745-3992.2005. 00002.x

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Chapman and Hall.

Gök, B., Kelecioğlu, H., & Dogan, N. (2010). The comparison of Mantel-Haenszel and logistic regression techniques in determining the differential item functioning. *Education and Science, 35*(156), 3-16.

Hambleton, R. K., Clauser, B. E., Mazor, K. M. & Jones, R. W. (1993). *Advances in the detection of differentially functioning test items*. University of Massachusetts, School of Education. http://files.eric.ed.gov/fulltext/ED356264.pdf

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the DIF analysis. *Test Validity*, 129-145.

Karasar, N. (2011*). Bilimsel araştırma yöntemleri*. Nobel.

Little, R., & Rubin, D. (2020). *Statistical analysis with missing data* (4th ed.). Wiley.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334. https://doi.org/10.1177/014662169301700401

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

Özgüven, İ. E. (2017). *Psikolojik testler*. Nobel.

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

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error Performance. *Journal of Educational Measurement, 33*(2), 215-230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Rousseau, M., Bertrand, R., & Boiteau, N. (2006). *Impact of missing data treatment on the efficiency of DIF methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Rubin D. B. (1976). *Inference and missing data*. *Biometrika*, *72*, 359-364.

Selvi, H., & Alıcı, D. (2018). Investigating the impact of missing data handling methods on the detection of differential item functioning. *International Journal of Assessment Tools in Education, 5*(1), 1-14. https://doi.org/10.21449/ijate.330885

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 on Measurement in Education, San Francisco, CA.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*(4), 545-571.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. https://doi.org/10.1037/1082-989X.7.2.147

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

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics* (3rd ed.). Herper Collins College Publishers.

Tamcı, P. (2018). *Kayıp veriyle başa çıkma yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi [Investigation of the impact of techniques of handling missing data on differential item functioning]* (Thesis No. 517260) [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Toka, O. (2012). *Kayıp veri durumunda sağlam kestirim [Robust estimation in case of missing data]* (Thesis No. 321449) [Master's thesis, Hacettepe University]. Council of Higher Education Thesis Center. https://tez.yok.gov.tr/UlusalTezMerkezi/

Turgut, M. F., & Baykul, Y. (2012). *Eğitimde ölçme ve değerlendirme []*. Pegem.

Van Buuren, S. (2012). *Flebitler imputation of missing data.* CRC Press.

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

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 232-243

Comparison of Methods of Affect Transition Analysis: An Example of SimInClass Dataset

Şeyma ÇAĞLAR ÖZHAN* Arif ALTUN**

Abstract

Studies of emotional-cognitive sequences are the growing body of research area in educational context. These studies focus on how emotions change during the learning-teaching process due to their dynamic nature. In affect transition studies, the change of emotion, depending on the event and time, is usually analyzed by using (a) lag sequential analysis (LSA), (b) L metric, (c) L* metric, and (d) Yule's Q metric. Yet, various methodological criticisms exist in the literature while utilizing these sequential analysis methods. In this study, it is aimed to comparatively examine lag analysis, L metric, L* metric, and Yule's Q in terms of proportion of invalid values, maximum transition metrics, minimum transition metrics, and analysis results. For this reason, the emotional states of the fifteen prospective teachers were collected and their emotions were labeled every 0.5 seconds with EEG (Electroencephalogram), GSR (Galvanic Skin Response), and Microsoft Kinect in a teacher training simulator (SimInClass). The dataset contained 17570 emotions, and the data were analyzed by utilizing lag analysis, L, L* and Yule's Q. The results showed that LSA had yielded the most proportion of invalid results. In addition, it was observed that the number of invalid values increased as the segment length became shorter in all analysis methods. When the maximum and minimum transition metric values were examined, it was found that as the sequence length increased in L and L* analyses, the value of the metrics approached 1, which is the largest value they can reach. However, it was noted that the lag analysis maximum-minimum transition metrics fluctuate independently from the sequence length. It was concluded that there were differences in the analysis results produced by the four sequential analysis methods with the same functions. It was thought that this situation might be due to the different invalid results produced by the analyses. When the results were compared with the studies in the literature, it was thought that it would be beneficial to pay attention to the nature of the data (emotional or behavioral), the data type (singe modality or multimodal modality), the amount of data (short sequences or long sequences), the environment in which the dataset was created (computer-based or not), and the sampling rate (automated data collection tool or observation) when choosing sequential analysis methods.

Keywords: affect, affect transition, sequential analysis

Introduction

Emotion is an intensely conscious mental response to a specific goal, subjectively experienced and lasting from minutes to hours, causing physiological and behavioral changes (Kleinginna & Kleinginna, 1981). The concept of emotion is sometimes referred to interchangeably with the terms "feeling," "mood," and "arousal" in the literature. The term "affect" is also used in a comparable way as emotion; it is known as a meta concept that covers emotion (Juslin & Slobada, 2013).

Emotions experienced by individuals in an educational context are the product of a cognitive evaluation structured by individuals' goals, teaching requirements, and competencies (Frenzel et al., 2009). According to appraisal theory, by updating any of the elements that construct emotions, individuals initiate the cognitive reappraisal process, and an affect transition occurs (transition from one emotion state to another) (Scherer,1993). In some studies, such as in Han et al., (2021) and Rebolledo-Mendez et al. (2022), affect transitions are examined to understand the effectiveness of intervention methods and the change of emotion in the learning-teaching process.

* Ress. Assist., Bartın University, Faculty of Sciences, Bartın-Turkey, seymacaglar9@gmail.com, ORCID ID: 0000-0002-0106-6285
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, altunar@hacettepe.edu.tr, ORCID ID: 0000-0003-4060-6157

To cite this article:
Çağlar Özhan, Ş., & Altun, A. (2022). Comparison of methods of affect transition analysis: An example of SimInClass dataset. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 232-243. https://doi.org/10.21031/epod.1051716

Received: 31.12.2021
Accepted: 12.09.2022

Instead of using simple ways to examine the affect transition, researchers apply sequential and conditional methods, among which are lag sequential analysis (LSA; Bakemann & Gottman, 1997), L metric (D'Mello & Graesser, 2012), L* metric (Matayoshi & Karumbaiah, 2020), and Yule's Q (Yule, 1900). In these methods, various measures such as likelihood or probability are used regarding the frequency of transition from one state to another (Bosch & Paquette, 2021). These methods are not only used for affect but also used in the discovery of patterns in computer-assisted environments, like exploring the behavioral patterns of learners at different success levels in a mobile learning environment (i.e. Sun et al., 2021), examining discussion orientations in online discussion environments according to cognitive styles (i.e. Wu & Hou, 2015), and examining development trajectories of learners' cognitive behaviors in small private online courses (i.e. Liu et al., 2021).

In studies where affect transition is explored, there are some methodological differences in examining sequential states. While in some studies self-transitions (temporal sequential repetition of an emotional state) are counted in the analysis, in others they are not. Some studies prefer Lag analysis when examining affect transition, while others prefer the L metric, L*, or Yule's Q.

For example, Lajoie et al. (2021) examined the emotional states of individuals according to their performance in a learning process that requires self-regulation. In each step of the self-regulated learning process, data on the mood of individuals and the temporal change of emotion were collected using a camera, and emotions were defined categorically by the software. For six states—happy, surprised, angry, scared, sad, and disgusted—analyses were conducted with 1786 emotional codes for the high-performance groups and 846 emotional codes for the low-performance group. The patterns in the sequences were examined with lag sequence analysis. Sequences do not include self-transitions in states. Lajoie et al.'s (2021) results showed that low-performing individuals set higher goals, experienced more negative emotions, and experienced more emotional transitions than high-performing individuals.

Baker et al. (2010) aimed to examine the persistence and occurrence rate of six emotional states, namely boredom, delight, confusion, concentration, frustration, and surprise, in three computer-assisted learning environments. Since the persistence of the emotion is also examined, self-transitions are included in the sequences. Analyses were performed on sequences of 706 and 3640 emotional codes using the L metric for two different implementations. They found that boredom had persistence in all three environments and was associated with poorer learning, while the affect state with the least persistence was frustration.

Botelho et al. (2018) aimed to explore the dynamics of students' affect in a computer-based learning environment. The affect states of the students were determined as sensor-free by using the behaviors in the system, and the dataset with 48276 observations was evaluated with the L metric analysis. Five different states, namely concentration, boredom, confusion, frustration, and neutral, took part in the participants' sequences. The results showed that the transitions from engagement to boredom and from confusion to engagement were significant.

Karumbaiah et al. (2021) aimed to examine methodological differences in the use of L metrics in affect transition studies. They emphasized that the number of affect states was a minimum of five (See. Botelho et al., 2018) and a maximum of thirteen (See. Bosch & D'Mello, 2017). The studies were carried out in computer-based tutoring systems and game-based learning environments, and in some of them, self-transition was excluded, while in others, it was included. In order to see the results of these differences, they examined ten different datasets used in previous studies and provided a corrected method for the use of L metrics. These investigations yielded that excluding the self-transition violated the assumption of independence causing more patterns to appear while providing no information on the persistence of affect states. They concluded that it is possible to obtain information about the permanence of the states even though the emotional transitions were diluted with the inclusion of self-transitions.

Bosch and Paquette (2021) examined different transition analyses from a methodological perspective. They compared the findings of sequential analysis methods by using two simulation datasets obtained in a computer-based environment. For this purpose, it was tested using simulated datasets containing 10,000 data. The comparisons were made by performing analyses with two and four variables using the dataset obtained from the simulation. They also analyzed another dataset with three variables: self-transitions were removed entirely, and the states showed a balanced distribution. They sequentially

examined 99 students' participation in practice and learning activities in a computer-based learning environment to compare the data produced by the simulation with data obtained from the natural environment. The sequence lengths of the participants ranged from 87 to 1087. In addition, analyses were applied to sequences with sequence lengths ranging from 5 to 50 extracted from these long sequences. The results showed that invalid values were higher for all sequential analysis metrics in short sequences than in the long sequences. In addition, invalid values increased as the number of variables increased in all sequential analyses. The maximum average transition value was calculated to determine whether all metrics deviated from zero. The results showed that short sequences produced artificially meaningful results in other transition metrics except for Yule's Q metric. To keep the invalid metrics under control, it has been stated that sequential analyses can be performed in sequences with at least ten observations for sequences with two variables; at least 20 for sequences with four variables; and, at least 35 observations for sequences with seven variables. Long sequences should be used to avoid false positive results.

In summary, transition metrics and related analysis methods are frequently preferred in studies exploring cognitive, behavioral, and affective patterns. However, differences in the applications of these analytical methods cause methodological concerns. In re-examining studies using L-metric analyses, Karumbaiah et al. (2021) found that excluding the states of repeating states in sequences and self-transition will result in discovering false patterns. Bosch and Paquette (2021), who examined the transition metrics and analyzed them comparatively, found that transition analysis methods produced different results according to the sequence length and number of affective states. However, Karumbaiah et al. (2021) focused only on the L metric, and Bosch and Paquette (2021) compared the analysis methods on the simulated data.

However, as Bosch and Paquette (2021) state, when examining behavior in online environments, there is no such thing as pressing a button many times in succession, so there is no repetitive data. When it comes to affect, the persistence state is high, and at the same time, the sampling rate is very high because of automatic emotion-recognition systems. These situations cause the formation of datasets with long sequences and repetitive data. In addition, due to the ever-changing and multidimensional nature of emotion, the number of states may be high in the data collected in natural environments; however, it may not be possible to distribute them in a certain balance. In this study, based on the findings and limitations in the literature, the goal is to perform and compare an analysis of transition metrics in data collected in the natural environment, where the seven affective states are included; the states are not evenly distributed, and self-transitions are included in the sequences. It is questionable, though, which sequential analysis method will work in the datasets with this feature and whether they will produce similar results.

From this point of view, the first aim of this study is to comparatively examine invalid values and maximum–minimum transition values produced by different methods, according to sequence length, by applying sequential analysis methods to the affect datasets formed by making sense of physical and physiological data in a virtual classroom simulation (SiminClass). The second aim is to compare the subsequent results by using the sequential analysis method.

As can be seen in the equations presented in the Methods section, the ratio is calculated with transition metrics. For this reason, when the denominator is equal to 0, the affected transition metrics cannot be calculated, and they produce an invalid value. That value is not calculated in all metrics for a sequence because their formulas are different. In this study, the proportion of invalid values in all values was examined among the analysis methods. Because the excess of invalid values in a sequence will reduce the probability of the occurrence of an expected value, the excess of these values reduces the statistical power (Bosch & Paquette, 2021). For this reason, it is important for the accuracy of the analysis to examine the status of invalid values according to the sequence length.

Another situation to be examined is the maximum–minimum values reached by the metrics. As mentioned in the Methods section, there are values in a certain range that each metric can tolerate. These values were also examined in order to understand how the trends of the metrics changed according to the sequence length. Each metric produced for each student's emotional transitions was evaluated within

itself, and maximum and minimum values were determined. In this context, the details of participants, the dataset, data collection tools, and data analysis are presented in the Methods section.

Method

Participants

Fifteen prospective teachers who took classroom management and teaching practice courses participated in the study. Prospective teachers were healthy individuals between the ages of 20-22, using their right hand dominantly. Before the application, the participants were informed through a consent form. Ethical permission for this study was obtained from the university.

Preparation of Dataset

The dataset in this study includes the emotional states of fifteen prospective teachers who teach in a virtual classroom simulation (SimInClass) during this process. In SimInClass, the participants plan the lesson, arrange the classroom seating, perform the teaching activities, and review the teaching report. Each session lasted approximately 6 to 10 minutes for each participant in the simulation. Simultaneously with the teaching process, physical and physiological data were collected from the participants. The participants' facial expressions were collected via Microsoft Kinect as physical data. As physiological data, the brain signals of the participants were obtained by EEG (Electroencephalogram), and galvanic skin responses were received by GSR. The data received from multiple modalities were interpreted by an emotion recognition system and labeled by the system to reflect a single dominantly experienced emotional state. Null values were observed when the physical and physiological sensors could not collect data. These values are not included in the dataset. A total of 17570 emotion codes were included in the dataset, in which labels were made according to seven basic emotional states. The minimum sequence length among the participants was 516, and the maximum sequence length was 1708. Details on data collection tools and emotion recognition are presented in the sub-headings.

Data Sources

In the study, physiological data were collected with GSR and EEG, and physical data were collected with Microsoft Kinect. With GSR, temperature and electrical changes in the skin with sweat and nerves are measured. Galvanic skin response occurs due to the difference in the electrical properties of the skin, which occurs in the interaction of individuals with stimuli (phasic) and the absence of any stimulation (tonic). This measure does not give a direct idea of emotions. It is related to the level of arousal of the sympathetic nervous system. It causes an activation similar to excitement, fear, anger, and surprise.

Electrical activation of the brain can be measured with EEG. It is a non-invasive measurement method that can obtain information about cognitive, affective, and motor functions by examining the activation created by the brain regarding a stimulus on time-base and frequency-base (Bayazıt, 2018). Microsoft Kinect makes sense of facial expressions with the facial mapping library.

Since it is essential to use multiple modalities in emotion recognition (Sebe et al., 2005) GSR, EEG, and facial expression data were interpreted with an emotion recognition module. The sampling rate of this module was approximately 0.5 sec. In the module, the data produced from GSR was combined with the data from facial expression by determining the arousal dimension of emotion, and the data produced from EEG determined the positive-negative dimension of emotion. As a result, the dominant one out of seven emotional states was determined by the value created. Three different (CNNA, CNNV, CNNF) InceptionResnetV2 Convolutional Neural Network (CNN) models were used to construct the multimodal emotion recognition model.

The model was tested on public datasets such as SEED and LUMED-LUMED2 datasets created by the developers of the emotion recognition module. Certain emotion categories were found to be confused

with others when the emotion recognition module identified emotions using only facial expressions. Emotions were determined with an accuracy rate of 74.2% when facial expression, EEG, and GSR were used together (see. Cimtay et al., 2020).

Data Analysis Method

The dataset was analyzed with four different sequential analysis methods and the findings were compared. The statistical significance of the sequential relationships between each emotion was examined via the L metric (D'Mello et al., 2007), L* metric (Matayoshi & Karumbaiah, 2020), lag sequential analysis-LSA (Bakeman & Gottman, 1997), and Yule's Q.

L metric is a transition likelihood metric that is used to determine affect transition and affect-behavior transition (D'Mello et al., 2007). The L metric indicates the probability and direction of the transition from the base frequency to the destination state. "Values above 0 signify that the transition is more likely than it could be expected to be given only the base frequency of the destination state, and values under 0 signify that the transition is less likely than it could be expected to be given only the base frequency of the destination state" (Baker et al., 2007). In this study, as suggested in the literature (D'Mello et al., 2007), D'Mello's L metric was calculated separately for each participant. One-sample t-test was applied to each participant's L metrics to test the transitions' significance. Benjamini and Yekutieli's (2001) post-hoc control method was applied to control false discoveries, as Matayoshi and Karumbaiah (2021) suggested. Transitions with an adjusted p-value less than .05 are considered statistically significant. The L metric formula is as follows:

$$L = \frac{P(Ynext|Xprev) - P(Ynext)}{1 - P(Ynext)}$$

L* is a transition metric modified from D'Mello's L metric. L* is a method of shifting the chance value of D'Mello's L from zero to a positive value to eliminate the undesirable effects of excluding self-transition (Matayoshi & Karumbaiah, 2020). L* is calculated using the formula for D'Mello's L metric. The only difference between the two methods is that when calculating the transition from A to B in L*, all transitions where A is the destination state are removed from the transition frequency matrix.

Lag Sequential Analysis (LSA) is an analysis method that is frequently used in the computer-based environment to examine the behavioral and affective interaction of users according to sequential and conditional probabilities (i.e Sun et al., 2017; Yang et al., 2018; Yang et al., 2018). In this study, as Bakeman and Gottman (1997) suggested, a coding scheme listing the emotions of prospective teachers in chronological order was prepared. Then, a transitional frequency matrix was created by calculating the frequency of each emotion category following another emotion category. In order to test the significance between the transitions, Z value was calculated using the transitional frequency. Z value represents the deviation of the probabilities of each transition from the expected values. It is accepted that Z values greater than +1.96 reach a significant level ($p < .05$). The Z value formula is as follows:

$$Z = \frac{f_{rc} - f_r p_c}{\sqrt{f_r p_c (1-p_c)(1-p_r)}}$$

Yule's Q is defined as a simple transformation of the odd ratio (OR) (Yule, 1900). Like the Pearson correlation coefficient, it takes a value between -1 and 1. Values between 0.3 and 0.7 indicate a moderate correlation between transitions, and values greater than 0.7 indicate a high correlation. In sequential analysis studies, it is mostly given with the Z value (i.e. Pohl et al., 2016). The Q metric formula is as follows:

$$Q = \frac{OR-1}{OR+1}$$

Results

1. Invalid Values and Maximum-Minimum Transition Values

In order to compare the results produced by the transition metrics in the real environment, the invalid values (i.e. "not a number values") and maximum-minimum transition values produced by the analyses were examined.

Figure 1 shows the proportion of invalid values for L, L*, and Z. In an affect dataset with sequence lengths ranging from 516 to 1708, it is seen that the proportion of invalid values decreases as the number of observations increases. In addition, the proportion of invalid values is higher in LSA (Z values) than in the L and L* metrics.

Figure 1

Invalid Values of Transition Metrics

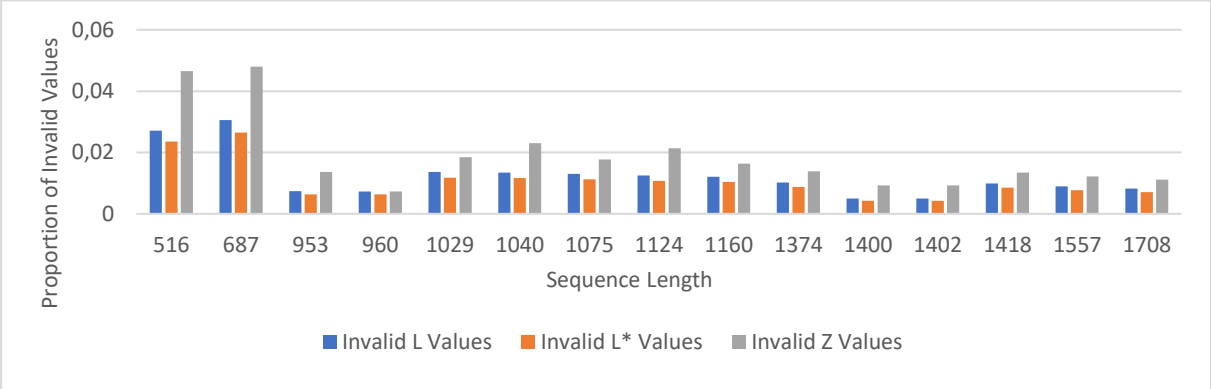
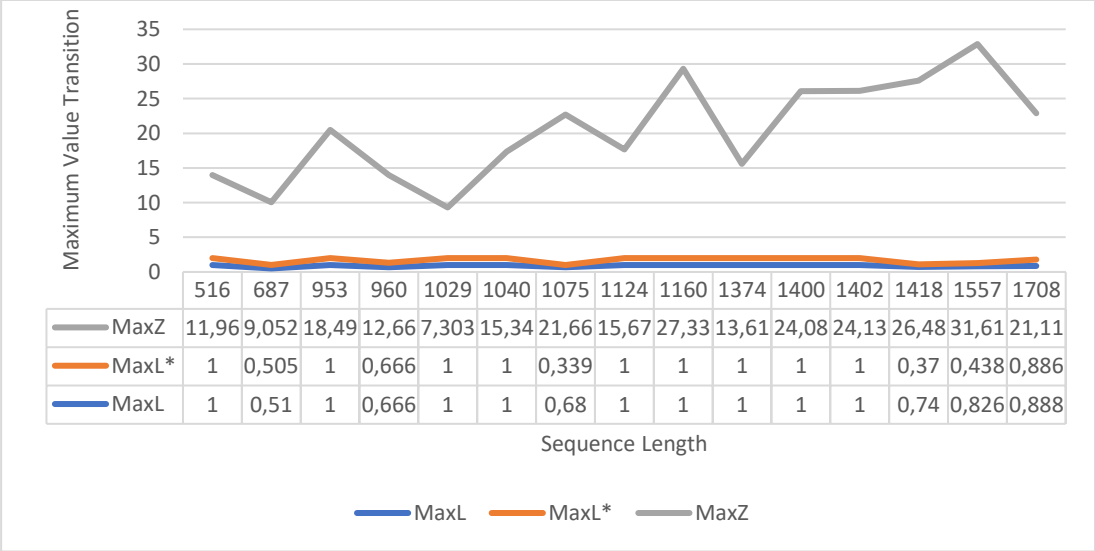


Figure 2a

Maximum Transition Values for L, L and Z values*



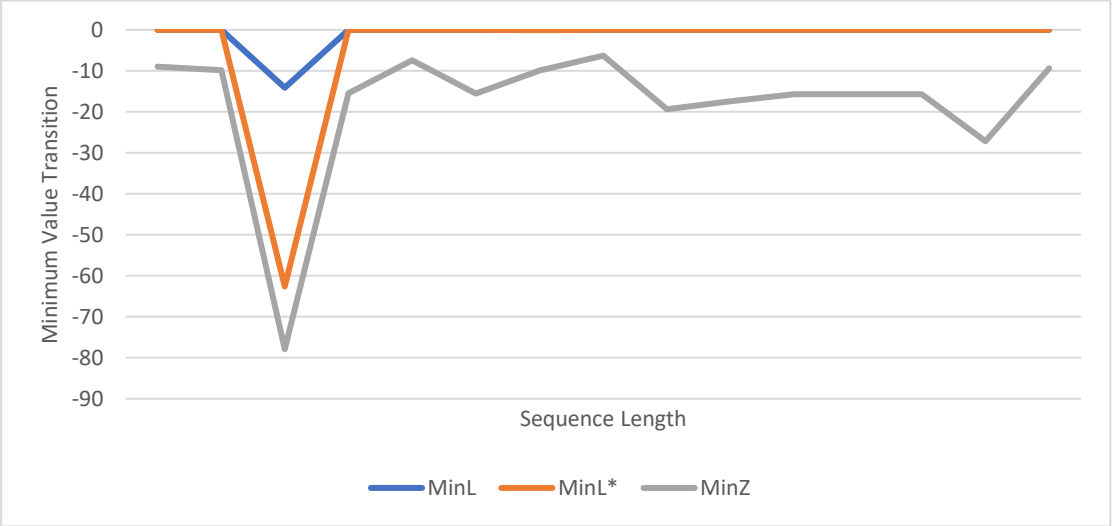
	516	687	953	960	1029	1040	1075	1124	1160	1374	1400	1402	1418	1557	1708
MaxZ	11,96	9,052	18,49	12,66	7,303	15,34	21,66	15,67	27,33	13,61	24,08	24,13	26,48	31,61	21,11
MaxL*	1	0,505	1	0,666	1	1	0,339	1	1	1	1	1	0,37	0,438	0,886
MaxL	1	0,51	1	0,666	1	1	0,68	1	1	1	1	1	0,74	0,826	0,888

In order to see how the trends of the metrics differ in an affect dataset collected in the real environment, maximum transition values (Figure 2a) and minimum transition values (Figure 2b) obtained from the analyses were examined. The results show that L and L* get closer to the maximum value of 1 as the

sequence length increases. In addition, it was observed that the minimum value of L and L* approached 0 for almost all transitions (except for one transition). Since the Z value in LSA can be between –∞ and ∞, it can reach no absolute maximum value. However, it was not observed that the maximum and minimum Z values produced from LSA increased proportionally with the sequence length. It is noteworthy that the maximum and minimum values of the Z value fluctuate independently of the sequence length.

Figure 2b

Minimum Transition Values for L, L and Z values*



2. Comparison of Affect Transition Analysis Results

In order to examine whether the affect transition analyses produce similar results in the SiminClass dataset, all analyses were applied to the dataset separately, and the results were examined comparatively. As seen in Table 1, the cyclical transition for happiness was statistically significant according to the L metric (Mean L = 0.64, SE = 0.03, t=19.6, p=0.00), LSA (ptr=0.89, z=103.53, p<.05) and Yule's Q (Q=0.97). Cyclical transition of sadness was also statistically significant according to the L metric (Mean L = 0.4, SE = 0.06, t=5.63, p=0.00), LSA (ptr=0.76, z=90.33, p<.05), and Yule's Q (Q=0.96). Cyclical transition for neutral was statistically significant according to the L metric (Mean L = 0.45, SE = 0.04, t=10.22, p=0.00), LSA (ptr =0.79, z=87.05, p<.05) and Yule's Q (Q=0.93). In addition, cyclical transition of disgust was statistically significant according to the L metric (Mean L = 0.29, SE = 0.07, t=4.2, p=0.001), LSA (ptr =0.55, z=67.81, p<.05) and Yule's Q (Q=0.96).

As seen in Table 1, the cyclical transition for anger was statistically significant according to LSA (ptr =0.17, z=15.9, p<.05) and Yule's Q (Q=0.99). According to LSA and Yule's Q, cyclical transition for surprise (ptr =0.33, z=39.81, p<.05; Q=0.97) and for fear (ptr =0.36, z=44.87, p<.05; Q=0.95) were statistically significant. In addition to cyclical transitions, the transition from fear to surprise (ptr =0.03; z=3.88, p<.05; Q=0.49), from surprise to fear (ptr =0.086; z=5.55, p<.05; Q=0.64), and from disgust to anger (ptr =0.004; z=2.85, p<.05; Q=0.61) are statistically significant for LSA and Yule's Q analyses.

Since self-transition is not included in L*, these findings would not inherently be common to L* as with other analyses. According to this analysis, the transition from happiness to disgust (Mean L*= 0.22, SE = 0.07, t=3.11, p<0.05) and from happiness to fear were meaningful (Mean L*= 0.26, SE = 0.08, t=3.12, p<0.05). Also, the transition from neutral to sadness (Mean L*= 0.18, SE = 0.05, t=3.15, p<0.05) and from neutral to disgust were also statistically significant (Mean L*= 0.18, SE = 0.64, t=2.85, p<0.05).

Table 1

Comparasion of Affect Transition Analysis

		Happiness	*Sadness*	*Neutral*	*Anger*	*Disgust*	*Surprise*	*Fear*
Happiness	**D'Mello L**	**0.64*** (0.03)	0.19 (0.06)	-0.72 (0.9)	-	0.24 (0.07)	-	0.27 (0.08)
	L*	-	**0.17*** (0.06)	-3.08 (3.24)	0.13 (0.09)	**0.22*** (0.07)	-	**0.26*** (0.08)
	LSA	**0.89***	0.007	0.07	0.00	0.02	0.003	0.012
	Yule's Q	**0.97***	-0.95	-0.87	-0.7	-0.49	0.51	-0.33
Sadness	**D'Mello L**	0.02 (0.01)	**0.4*** (0.06)	-0.82 (0.91)	-	0.06 (0.02)	-	0.11 (0.06)
	L*	0.02 (0.01)	-	-0.99 (1.08)	0.05 (0.04)	0.06 (0.02)	-	0.11 (0.06)
	LSA	0.02	**0.76***	0.18	0.001	0.03	0.001	0.08
	Yule's Q	-0.94	**0.96***	-0.43	0.009	-0.3	-0.73	-0.43
Neutral	**D'Mello L**	0.11 (0.03)	0.2 (0.05)	**0.45*** **(0.04)**	0.95 (0.07)	0.18 (0.06)	-	0.16 (0.06)
	L*	-0.02 (0.08)	**0.18*** (0.05)	-	0.09 (0.07)	**0.18*** (0.64)	-	0.16 (0.06)
	LSA	0.07	0.09	**0.79***	0.001	0.02	0.01	0.011
	Yule's Q	-0.88	-.045	**0.93***	-0.009	-0.55	0.1	-0.3
Anger	**D'Mello L**	-0.002 (0.003)	0.03 (0.02)	-0.934 (0.9)	0.09 (0.07)	0.02 (0.016)	-	-0.0001 (0.0001)
	L*	-0.002 (0.003)	0.03 (0.02)	-0.99 (0.9)	-	0.02 (0.01)	-	-0.0001 (0.0001)
	LSA	0.09	0.09	0.45	**0.27***	0.09	0.00	0.00
	Yule's Q	-0.31	0.009	-0.14	**0.99***	0.1	-1	-1
Disgust	**D'Mello L**	0.17 0.01	0.02 0.008	-0.91 0.91	0.15 (0.09)	**0.29*** (0.07)	-	0.03 (0.02)
	L*	0.01 (0.01)	0.019 (0.009)	-0.97 (1.00)	0.15 (0.009)	-	-	0.03 (0.02)
	LSA	0.18	0.09	0.13	**0.004***	**0.55***	0.004	0.024
	Yule's Q	-0.51	-0.31	-0.53	**0.61***	**0.96***	-0.22	0.06
Suprise	**D'Mello L**	-0.00 (0.004)	0.007 (0.007)	-0.92 (0.91)	-	0.002 (0.002)	-	0.018 (0.009)
	L*	-0.0003 (0.004)	0.007 (0.008)	-0.92 (0.94)	-0.00 (0.00)	0.002 (0.002)	-0.00 (0.00)	0.018 (0.009)
	LSA	0.15	0.02	0.40	0.00	0.01	**0.33***	**0.09***
	Yule's Q	-0.47	-0.85	0.08	-1	-0.73	**0.97***	**0.64***
Fear	**D'Mello L**	0.006 (0.005)	0.057 (0.032)	-0.91 (0.91)	-.0003 (0.000)	.022 (0.01)	-	0.16 (0.05)
	L*	0.004 (0.005)	0.057 (0.033)	-0.94 (0.97)	-.0003 (0.000)	.022 (0.01)	-	-
	LSA	0.29	0.07	0.20	0.00	0.05	**0.03***	**0.36***
	Yule's Q	-0.31	-0.36	-0.33	-1	0.03	**0.49***	**0.95***

* Statistically significant transitions are shown in bold.

**Mean L and Mean Standard Error for D'Mello L method, Mean L* and Mean Standard Error for L* method, and Z value for LSA are given in the table.

Discussion and Conclusion

In this study, it was aimed to comparatively examine the invalid values and maximum-minimum transition values produced by different sequential analysis methods in the affect dataset, which was formed by making sense of physical and physiological data in a virtual classroom simulation. In addition, the results produced by the sequential analysis methods were examined comparatively and

based on the theoretical framework. In this context, an implementation was carried out with fifteen prospective teachers in a classroom simulation (SiminClass) to create a dataset. Participants' affect states were recognized every 0.5 seconds using EEG, GSR, and facial expressions to include seven basic emotional states. The sequential analysis methods L, L*, LSA, and Yule's Q were applied to the created dataset.

After applying the sequential analysis methods, the invalid values produced by L, L*, and LSA were compared. The results showed that the proportion of invalid values of LSA is higher than L and L*. Moreover, as the sequences get shorter, invalid values increase in all analysis methods. As a matter of fact, in the study of Bosch and Paquette (2021), it was found that LSA produced more invalid values in the log dataset containing behavioral data compared to other sequential analysis methods, and short sequences produced more invalid values than long sequences. Due to invalid values, the number of transitions that show a meaningful pattern in a dataset may decrease, and the transition parameters may be found less than they should be. This issue increases the chance of finding outliers (Bosch & Paquette, 2021). This may lead to type 1 or type 2 error. Therefore, analysis can be made using L or L* methods in data collected in the real environment, where there are seven states and the distribution ratio of the states is random. In addition, it is predicted that the sequences being as long as possible will increase the statistical power of the results obtained from the sequential analysis method to be used.

When the maximum and minimum values obtained from the analyses were examined, considering the limits of L and L* [–∞,1], it was seen that the maximum value could be approached as the sequence length reaches a certain saturation in both analyses. It has been observed that the minimum value curve converges at zero for L and L*. This indicates that at least one transition between states never occurred. Since it was a dataset created in the real environment, the transitions between the states were random; therefore, this result was expected (Matayoshi & Karumbaiah, 2020). There is no maximum or minimum value that the Z value produced with LSA can take. However, the findings show that the maximum and minimum values of LSA fluctuate independently of the number of observations.

On the contrary, Bosch and Paquette (2021) found that as the number of observations increased, the maximum value of LSA increased, and the minimum value decreased. This difference may be due to the characteristics of the datasets. Bosch and Paquette (2021) performed LSA with sequence lengths ranging from 5 to 50 and log data in a computer-based environment. This finding could not be reached in the dataset with seven affect states and a sampling rate of 0.5 seconds. This shows that the nature of the data (emotional or behavioral), the data type (singe modality or multimodal modality), the amount of data (short sequences or long sequences), the environment in which the dataset was created (computer-based or not), and the sampling rate due to the characteristics of the data collection tool (automated data collection tool or observation) are essential. In sequential analyses, the analysis method to be used can be decided by paying attention to these dataset features.

When the analysis results are examined, the cyclical transition of happiness, sadness, neutral, and disgust is statistically significant according to the results of L, LSA, and Yule's Q. However, anger, surprise, and fear's cyclical transition were statistically significant only according to LSA and Yule's Q metric (cyclical transitions were not calculated since self-transitions were not examined in L*). When the results produced by the analyses are examined, the cyclical transition of happiness, sadness, neutral, and disgust is statistically significant according to the results of L, LSA, and Yule's Q. However, anger, surprise, and fear's cyclical transition were statistically significant only according to LSA and Yule's Q metric (cyclical transitions were not calculated since self-transitions were not examined in L*).

When these results were evaluated in the theoretical framework, it was an expected result that the cyclical transition of more intense and long-lasting emotions such as sadness would be found to be statistically significant. The statistical significance of the surprise cyclical transition did not match the theoretical framework. As surprise is the shortest of all emotions, it lasts for a few seconds and disappears (Ekman, 2021). For this reason, the statistical significance of the surprise cyclical transition did not match the theoretical framework.

The statistical significance of transitions between different emotions differed between analysis methods. According to LSA, the mutual pattern between fear and surprise is statistically significant, and there is

a moderate relationship between these transitions according to Yule's Q. According to LSA, the pattern of transition from disgust to anger was also statistically significant, and the Yule's Q metric shows a moderate relationship between these two states. However, there was no statistically significant transition between these states according to D'Mello's L and L*. This situation among the results of the analysis may have caused these contradictory results in the dataset, as more invalid values were produced in the LSA and the rare occurrence of emotions such as anger and fear.

Since state persistence is ignored in L* sequences, the results it naturally produces are different from the others. This metric can be used in studies that focus only on transitions between states where state persistence is not essential (Karumbaiah et al., 2019).

Studies examining emotional-cognitive sequences in educational contexts are a growing body of research area. However, it has been observed that different results were obtained for the same dataset in different methods of affect transition analysis serving the same purpose. To increase statistical power in studies examining affect transition, it is important to carefully examine the structure of the dataset and the purpose of the research. In this way, the appropriate affect transition analysis method can be selected. The data type, amount of data, the environment in which the data was created, and the characteristics of the data collection tool can be the determinants of this decision process. Also, in similar studies, it would be helpful to have the sequences as long as possible to avoid potential errors. On the other hand, should a transition frequency, which is very rare in long sequences, be excluded? Should the number of observations be determined according to the number of states in such studies? It is one of the topics recommended to be examined in future studies. In addition, this study was conducted in a classroom simulation with a limited number of participants and an automatic emotion identification system with little margin of error. The research can be conducted with more participants in different contexts and data collection tools in future studies.

Declarations

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

Ethical Approval: This study was approved by the Ethics Boards and Commissions of Hacettepe University (date 22.05.2020, document number GO20/459).

Acknowledgments: This work was supported by the Newton-Katip Çelebi fund, delivered by the Scientific and Technological Research Council of Turkey (TÜBİTAK) through the project titled ‘Investigating the Effects of Computer-Based Affective Recommendation System on Teacher Trainees Cognitive-Emotional Development' (Grant No:117R036).

Author Contribution: Caglar-Ozhan S. and Altun A. contributed to the design and implementation of the research to the analysis of the results. All authors discussed the results and commented on the manuscript.

Studies included in the current reliability generalization meta-analysis are marked with an asterisk (*).

References

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge University Press. https://doi.org/10.1017/CBO9780511527685

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223-241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Baker, R. S., Rodrigo, M. M. T., & Xolocotzin, U. E. (2007). The dynamics of affective transitions in simulation problem-solving environments. In A. Paiva, R. Prada & W. Picard (Eds.), *International conference on affective computing and intelligent interaction* (pp. 666-677). Springer. https://doi.org/10.1007/978-3-540-74889-2_58

Bayazıt, T. (2018). Event Related Potentials (ERP). *Journal of Medical Clinics, 1*(1), 59-65. https://dergipark.org.tr/tr/pub/atk/issue/38771/451155

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics, 29*(4),1165-1188. https://doi.org/10.1214/aos/1013699998

Bosch, N., & D'Mello, S. (2017). The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education, 27*(1), 181-206. https://doi.org/10.1007/s40593-015-0069-5

Bosch, N., & Paquette, L. (2021). What's next? Sequence length and impossible loops in state transition measurement. *Journal of Educational Data Mining, 13*(1), 1-23. https://eric.ed.gov/?id=EJ1320638

Botelho, A. F., Baker, R., Ocumpaugh, J., & Heffernan, N. (2018). Studying affect dynamics and chronometry using sensor-free detectors. In E. Boyer & M. Yudelson (Eds.), *Proceedings of the 11th international conference on educational data mining* (pp. 157–166). EDM. https://files.eric.ed.gov/fulltext/ED593106.pdf

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. (2020). Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access*, *8*, 168865-168878. https://doi.org/10.1109/ACCESS.2020.3023871

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145-157. https://doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S., Taylor, R., & Graesser, A. (2007). Monitoring affective trajectories during complex learning. In McNamara, D., Trafton, J. (Eds.), *Proceedings of 29th annual cognitive science society* (pp. 203–208). Cognitive Science Society. https://doi.org/10.1007/978-1-4419-1428-6_849

Ekman (2021, September 7). *What is surprise?* Paul Ekman. https://www.paulekman.com/universal-emotions/what-is-surprise/

Frenzel, A. C., Goetz, T., Stephens, E. J., & Jacob, B. (2009). Antecedents and effects of teachers' emotional experiences: An integrated perspective and empirical test. In P. A. Schutz & M. Zembylas (Eds.), *Advances in teacher emotion research: The impact on teachers' lives* (pp. 129-152). Springer. https://doi.org/10.1007/978-1-4419-0564-2_7

Han, J.-H., Shubeck, K., Shi, G.-H., Hu, X.-E., Yang, L., Wang, L.-J., Zhao, W., Jiang, Q., & Biswas, G. (2021). Teachable agent improves affect regulation: Evidence from Betty's brain. *Educational Technology & Society, 24*(3), 194–209. https://www.jstor.org/stable/27032865

Juslin, P. N., & Sloboda, J. A. (2013). Music and emotion. In D. Deutsch (Ed.), *The psychology of music* (pp. 583-645). Academic Press. https://doi.org/10.1016/B978-0-12-381460-9.00015-8

Karumbaiah, S., Baker, R. S., & Ocumpaugh, J. (2019). The case of self-transitions in affective dynamics. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *International conference on artificial intelligence in education* (pp. 172-181). Springer. https://doi.org/10.1007/978-3-030-23204-7_15

Karumbaiah, S., Baker, R. B., Ocumpaugh, J., & Andres, A. (2021). A re-analysis and synthesis of data on affect dynamics in learning. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2021.3086118

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5*(4), 345-379. https://doi.org/10.1007/BF00992553

Lajoie, S. P., Zheng, J., Li, S., Jarrell, A., & Gube, M. (2021). Examining the interplay of affect and self regulation in the context of clinical reasoning. *Learning and Instruction, 72*, 101219. https://doi.org/10.1016/j.learninstruc.2019.101219

Liu, Z., Zhang, N., Liu, S., & Liu, S. (2021). Development trajectory of student cognitive behaviors in a SPOC forum: An integrated approach combining epistemic network analysis and lag sequential analysis. In L. Lee, L. Wang, Y. Kato & S. Sato (Eds.), *2021 International symposium on educational technology (ISET)* (pp. 26-30). IEEE. https://doi.org/10.1109/ISET52350.2021.00016.

Matayoshi, J., & Karumbaiah, S. (2020). Adjusting the L statistic when self-transitions are excluded in affect dynamics. *Journal of Educational Data Mining, 12*(4), 1-23. https://eric.ed.gov/?id=EJ1298368

Matayoshi, J., & Karumbaiah, S. (2021). Using marginal models to adjust for statistical bias in the analysis of state transitions. In M. Scheffel, N. Dowell, S. Joksimovic & G. Siemens (Eds.), *LAK21: 11th International learning analytics and knowledge conference* (pp. 449-455). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448182

Rebolledo-Mendez, G., Huerta-Pacheco, N. S., Baker, R. S., & du Boulay, B. (2022). Meta-affective behaviour within an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education, 32*(1), 174-195. https://doi.org/10.1007/s40593-021-00247-1

Pohl, M., Wallner, G., & Kriglstein, S. (2016). Using lag-sequential analysis for understanding interaction sequences in visualizations. *International Journal of Human-Computer Studies, 96*, 54-66. https://doi.org/10.1016/j.ijhcs.2016.07.006

Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion, 7*(3), 325–355. https://doi.org/10.1080/02699939308409192

Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. In C. Chen & P. Wang (Eds.), *Handbook of pattern recognition and computer vision* (pp. 387-409). World Scientific. https://doi.org/10.1142/1802

Sun, J. C. Y., Kuo, C. Y., Hou, H. T., & Lin, Y. Y. (2017). Exploring learners' sequential behavioral patterns, flow experience, and learning performance in an anti-phishing educational game. *Journal of Educational Technology & Society, 20*(1), 10-20. https://www.proquest.com/scholarly-journals/exploring-learners-sequential-behavioral-patterns/docview/2147743221/se-2

Sun, Z., Lin, C. H., Lv, K., & Song, J. (2021). Knowledge-construction behaviors in a mobile learning environment: A lag-sequential analysis of group differences. *Educational Technology Research and Development, 69*(2), 533-551. https://doi.org/10.1007/s11423-021-09938-x

Wu, S. Y., & Hou, H. T. (2015). How cognitive styles affect the learning behaviors of online problem-solving based discussion activity: A lag sequential analysis. *Journal of Educational Computing Research, 52*(2), 277-298. https://eric.ed.gov/?id=EJ1076314

Yang, X., Song, S., Zhao, X., & Yu, S. (2018). Understanding user behavioral patterns in open knowledge communities. *Interactive Learning Environments, 26*(2), 245-255. https://doi.org/10.1080/10494820.2017.1303518

Yang, X., Li, J., & Xing, B. (2018). Behavioral patterns of knowledge construction in online cooperative translation activities. *The Internet and Higher Education, 36*, 13-21. https://doi.org/10.1016/j.iheduc.2017.08.003

Yule, G. U. (1900). On the association of attributes in statistics: With illustrations from the material of the childhood society. *Philosophical Transactions of the Royal Society of London. 66*(194), 252–261. https://doi.org/10.1098/rspl.1899.0067

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 244-255

Bifactor and Bifactor S-1 Model Estimations with Non-Reverse-Coded Data

Fulya BARIŞ PEKMEZCİ*

Abstract

The bifactor model is an extension of Spearman's two-factor theory. The bifactor model has a strict assumption, which is named orthogonality. The bifactor S-1 model was developed by stretching the orthogonality assumption of the bifactor model. The bifactor S-1 model, contrary to the bifactor model, allows correlation between specific factors and enables items that do not form a common specific factor to be loaded only on the general factor. In psychology, data are mostly multidimensional due to the nature of psychological constructs. The Positive and Negative Affect Schedule (PANAS) which is one of the psychological tests and has two dimensions named positive affect and negative affect. In the literature studies on PANAS, negative affect dimensions were not reverse coded while implementing the bifactor model. Therefore, negative path coefficients were revealed. The purpose of this study is to ascertain whether or not the items in the negative affect factor should be reverse coded in the PANAS. Within the scope of the current study, bifactor and bifactor S-1 model analyses were implemented for the two data sets, which were reverse coded and non-reverse coded. As a result of this study, with reverse-coded data, the bifactor S-1 model was seen as the better model for the PANAS. Additionally, in the modeling of unique variances of items with specific factors, the bifactor S-1 model performed well and also resolved the problem of negative loading on the general factor. The point to take into consideration, which should be noted by researchers who will study the PANAS, is that negative items should be reverse coded.

Keywords: PANAS, bifactor S-1, reverse coding

Introduction

The bifactor model and methods were introduced by Holzinger and Swineford (1937) as an extension of Spearman's two-factor theory. According to Spearman's two-factor (g-factor) theory, individuals have a single cognitive capacity, named "g". The "g", which does not change throughout life, consists of abstract thinking and problem-solving skills, and the ability to perform complex mental processes. The factor named "s", consists of the individual's specific skills concerned with mathematical and verbal ability. According to Spearman's conceptualization of cognitive skills, all variables are related to a general factor and at least one specific factor. Figure 1 shows the bifactor model that items can be loaded on two different factors, named the general and specific factors (Canivez, 2016; Reise et al., 2010). All items loaded on general factor and items shared common content are loaded at the same specific factor. The bifactor model, which is used to separate the contributions of specific facets/factors to the general factor, is frequently used for scaling psychological constructs. In addition, the bifactor model is used to create a short unidimensional scale from a multidimensional or unidimensional scale (Stucky & Edelen, 2015; Stucky et al., 2014) and is useful in terms of using subscale scores (Cucina & Byle, 2017). Besides these advantages, in addition to Item Response Theory (IRT) assumptions, the bifactor model has a strict assumption, which is named orthogonality. The orthogonality assumption requires that the specific factors are orthogonal to each other and the general factor, in other words, there is no correlation between these factors. However, the bifactor S-1 model (Figure 1) was developed as a result of stretching the orthogonality assumption of the bifactor model. Contrary to the bifactor model, the bifactor S-1 model enables correlation between specific factors. Besides, the bifactor S-1 model enables items that do not form a common specific factor to

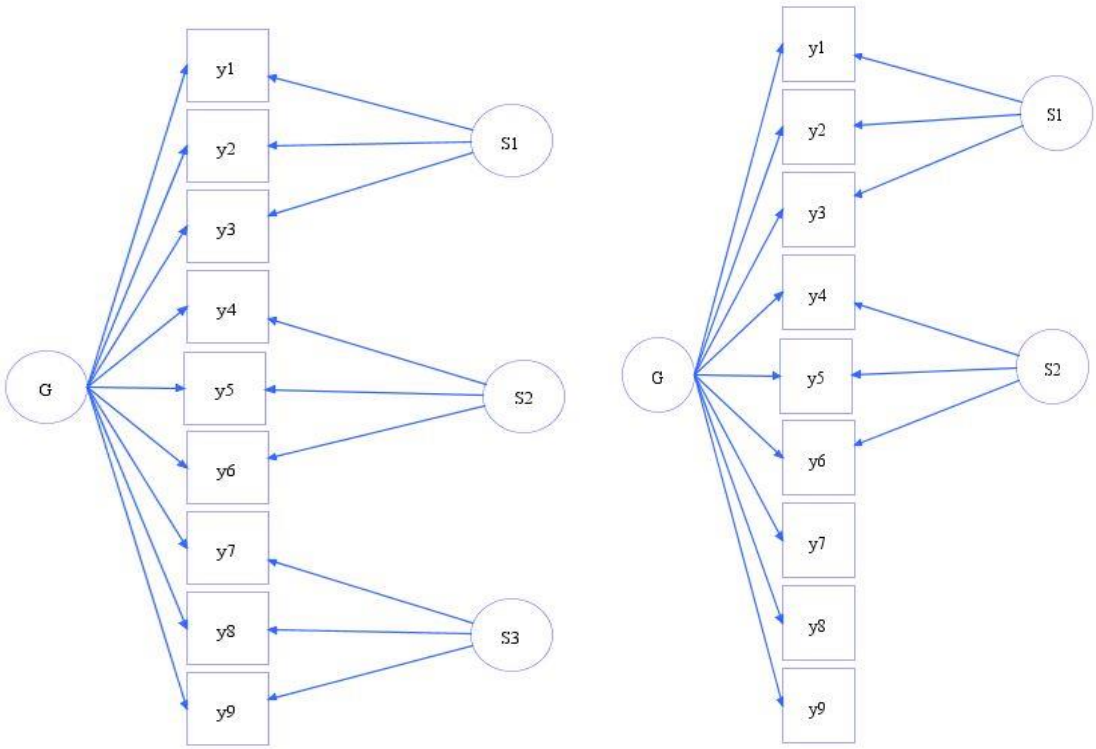
* Assoc. Prof. Dr., Yozgat Bozok University, Faculty of Education, Yozgat-Turkey, fulya.baris@bozok.edu.tr, ORCID ID: 0000-0001-6989-512X

To cite this article:
Baris-Pekmezci, F., (2022). Bifactor and bifactor S-1 model estimations with non-reverse-coded data. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 244-255. https://doi.org/10.21031/epod.1135567
Received: 24.06.2022
Accepted: 26.09.2022

load only on the general factor and can be applied to many psychological constructs (Burns et al., 2020; Eid, 2020).

Figure 1

Example of Bifactor and Bifactor S-1 model



G: General factor; S: Specific factor; y: items

In psychology, data obtained from any constructs are mostly multidimensional, and these dimensions correlated with each other. The Positive and Negative Affect Schedule (PANAS) is one of the scales in psychology, developed by Watson et al. (1988). It has two dimensions, named *positive affect* (PA) and *negative affect* (NA). Watson et al. (1988) proved that two factors are orthogonal. Since then, the PANAS has been developed and is widely used with clinical and non-clinical samples (Flores-Kanter et al., 2021; Rush & Hofer, 2014). However, the construct validity of the PANAS has been debated for almost 30 years (e.g., Ebesutani et al., 2011; Gaudreau et al., 2006; Watson et al., 1988). Also, there are various studies that have different interpretations and findings about the factor structure of the PANAS (Huebner & Dew, 1995; Killgore, 2000; Leue & Beauducel, 2011; Mihić et al., 2014; Ortuño-Sierra et al., 2015; Pires et al., 2013; Seib-Pfeifer et al., 2017; Vera-Villarroel et al., 2017). Besides, the PANAS factor structure is more suitable for bifactor model analysis due to its consisting of two orthogonal factors, which are PA and NA. In the literature, studies that modelled PANAS according to the bifactor model used naming affective *polarity* (AF) for general factor (Leue & Beauducel, 2011), and in this study, the same terminology was used.

In studies (Leue & Beauducel, 2011; Seib-Pfeifer et al., 2017) that used a bifactor model to examine the PANAS factor structure, negative path coefficients were revealed due to non-reverse coding of negative items. However, Brown and Marshall (2001) and Zampetakis et al. (2015) stated that those who work with PANAS should reverse code the items regarding the NA factor.

According to DiStefano et al. (2009), an item with a negative factor loading means a negative correlation between the item and factor, and the raw score of the item should be subtracted rather than added. From the factor analytical perspective, in exploratory factor analysis (EFA), factor scores mean correlation coefficients between the item and factor (Bernard, 2013; Kline, 2005), while for confirmatory factor analysis (CFA), factor scores mean path coefficients.

Negative factor loadings show that the items measure the opposite trait of the determined factor. At this point, the general factor loses its importance for the bifactor model. Although the general factor consists of all scale items, the negative correlation between an item and the general factor shows that the specific item measures a different trait from the general factor. So, the debate arises as to whether the negatively correlated items should be included in the general factor. According to DiStefano et al. (2009), negatively correlated items should be subtracted from the related factor. This means that negatively correlated items will not contribute to the explained variance in the general factor. In this situation, the existence of the general factor becomes questionable. If negative items are not reverse coded, can the bifactor model be used?

From this point of view, the purpose of this study is to ascertain whether or not the items in the NA factor should be reverse coded in the PANAS while using bifactor models. With this purpose in mind, the PANAS items were modeled with both the bifactor and bifactor S-1. In the literature, there were two studies that the PANAS were modelled according to the bifactor model. To compare the results with the literature in addition to bifactor s-1, bifactor model analyses were implemented too. The Bifactor s-1 model was used because it is not required an orthogonality assumption and enables avoiding estimation bias arising from correlation. It is hoped that this research will contribute to a deeper understanding of reverse coding of negative items. In line with this purpose, the following research questions were asked:

1) What are the factor loadings and model fit statistics according to the bifactor model when the items in the NA factor are not reverse coded?
2) What are the factor loadings and model fit statistics according to the bifactor model when the items in the NA factor are reverse coded?
3) What are the factor loadings and model fit statistics according to the bifactor S-1 model when the items in the NA factor are not reverse coded?
4) What are the factor loadings and model fit statistics according to the bifactor S-1 model when the items in the NA factor are reverse coded?

Methods

Participants

In this study, data were obtained from the Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel (www. lissdata.nl). The survey research data of the LISS Core Study of Personality Wave 11 were used in this study. The sample of Wave 11 consisted of 5075 participants aged 16 or older.

Instrument

The PANAS is an adjective-based scale which has 10 items for NA and 10 items for PA (Magyar-Moe, 2009). Participants indicate "*To what extent do you feel in general?*" (1=*not at all*, 5=*extremely*) for each item. The whole data set had 5201 cases. In structural equation modelling, there is always the risk of accepting invalid models with very "large" sample sizes and of rejecting valid models with very "small" sample sizes (Molwus et al., 2013). Kline (2011) stated that for more complicated models sample size is at least 200 and Hair et al.

(2008) suggest a minimum of 200 and a maximum of 400 as appropriate sample size. According to Lacobucci (2010), 50 can be sufficient for minimum sample size and 100 can be sufficient for maximum sample size, and the rules of thumb suggesting required sample sizes to be at least 200 are "conservative" and "simplistic. To eliminate the large sample effect on the model fit indices, whole data were not used and two samples which have 1000 cases were drawn. Before bifactor and bifactor s-1 analysis, outliers were detected and deleted from the data. The sample size was 572 for reverse-coded data and 732 for non-reverse-coded data. Final samples have differences in sample sizes due to the number of outliers.

Analysis

Before starting the analysis, two samples have 1000 cases were drawn from downloaded data (n=5210). Then, data were recoded to obtain reverse-coded data for the bifactor model and bifactor s-1 model analyses. Then, both reverse-coded and non-reverse-coded data were cleaned for missing values and outliers to meet the assumptions of confirmatory factor analysis models (Kline, 2011). All data analyses were carried out with Mplus 7 software.

Some goodness-of-fit indices for confirmatory models were used to decide the best model fit. The fit indices and reference ranges which were obtained from Mplus 7 and used in this study are given in Table 1.

Table 1

Goodness-of-fit indices for confirmatory models

	Acceptable fit	Good fit
Chi-square (X^2)[3]	$2df<X^2\leq 3df$	$0\leq X^2\leq 2df$
Standardized root mean square residual (*SRMR*)[1]	$0.05<SRMR\leq 0.10$	$0.00\leq SRMR\leq 0.05$
The comparative fit index (*CFI*)[1]	$0.90\leq CFI<0.95$	$0.95\leq CFI\leq 1.00$
Tucker Lewis fit index (*TLI*)[1]	$0.90\leq TLI<0.95$	$0.95\leq TLI\leq 1.00$
Root mean square error of approximation (*RMSEA*)[2]	$0.05<$RMSEA≤ 0.08	$0.00\leq$RMSEA≤ 0.05

Note: [1]For SRMR, CFI, and TLI, cutoff values were obtained from Hu and Bentler (1999) and Schermelleh-Engel, Moosbrugger & Müller (2003; p.52)

[2]For RMSEA, cutoff values were obtained from Brown (2015).

[3]For, $\mathcal{X}^2$ Schermelleh-Engel, Moosbrugger & Müller (2003).

For $\mathcal{X}^2$, insignificant values indicate that the model-data fit is provided. However, $\mathcal{X}^2$, is always affected by the sample size (Distefano & Hess, 2005; Kline, 2011). With large samples, $\mathcal{X}^2$ tends to be significant (Zimmer & Odum Institute, 2019).

In the scope of this study, firstly bifactor analysis was performed for both non-reverse coded data and reverse-coded data. Bifactor analysis results were used to determine which items (Positive or negative items) were to be used for loading in the AF factor.

Results

As aforementioned, two data sets were prepared which are reverse coded and non-reverse coded data. Bifactor analysis was performed for both non-reverse coded data and reverse-coded data for PANAS. The results of the bifactor model analysis are given in Table 2.

Table 2

Bifactor analysis for reverse coded and non-reverse coded data

Coding type	Models	RMSEA	RMSEA CI		CFI	TLI	SRMR	Chi-Square
non-reverse coded data	Without modification	0.116	0.111	0.121	0.875	0.843	0.397	1639.831 (151)
	With modification	0.104	0.099	0.110	0.900	0.873	0.397	1337.046 (149)
Reverse coded data	Without modification	0.099	0.093	0.105	0.910	0.887	0.367	1001.044 (151)
	With modification	0.085	0.080	0.091	0.934	0.916	0.368	771.100 (149)

When the model fit statistics in Table 2 were examined, it was seen that the model data fit was not achieved. To improve model fit, correlation of error terms can be used. Therefore, modifications were made through error term correlations. Error term modifications show that there is some common issue, which is not specified within the model, which causes covariation (Gerbing & Anderson, 1984). From this point of view, to provide model data fit, certain modifications (correlations of item PA9 with item PA8, and item PA9 with item PA6 in the PA factor were allowed) were made. Nevertheless, the model data fit was not achieved. Additionally, one of the modifications was remarkable, which suggests correlation between PA and AF (general factor) of the model. Besides, the bifactor model is rigid for factor correlations and does not allow correlation between general and specific factors. Thus, the fact that this remarkable modification could not be implemented. However, it can be evaluated as a preliminary finding for the compatibility of the bifactor S-1 model for the PANAS.

After obtaining the non-fitting model for the non-reverse data, bifactor analysis was performed for the reverse-coded data. As seen in Table 2, like with the non-reverse data, model data fit was not achieved. Some modifications (correlations of item PA4 with item PA5, and of item PA9 with item PA6 in the PA factor were allowed) were made to ensure model-data fit. However, the model data fit could not be achieved and it was observed that items NA3, NA4, and NA5 gave negative path coefficients with the NA factor. In addition to this, as a result of reverse coding, item PA2 resulted in a negative path coefficient with the general factor. In accordance with this result, previous studies demonstrated that item PA2 had a negative factor loading (Leue & Beauducel, 2011; Seib-Pfeifer et al., 2017). As with the data that were non-reverse coded, the correlation between PA and AF was seen as a remarkable modification for the model. Also, it is an important finding that the reverse-coded data showed a better fit than the non-reverse data. By obtaining the same results as for the non-reverse data, the preliminary finding was strengthened for the suitableness of the bifactor S-1 model.

When the factor loadings were analyzed according to bifactor analysis for the non-reverse data (see Appendix for detailed information), the factor loadings of four items in the NA factor and one item in the PA factor were lower than 0.32 (Comrey & Lee, 1992). Lower factor loadings in specific factors (PA and NA) were an indication of inadequate variance explained by the items. For reverse-coded data, only PA2 had a negative factor loading on the general factor (AF). Except for two items, all items in the NA factor had factor loadings lower than 0.32. Besides, three items had negative factor loadings in the NA factor. Modifications did not change the results.

As a result of obtaining the same finding for non-reverse and reverse-coded data, which showed a strong correlation between the PA factor and general factor (AF) and gave a remarkable modification coefficient (Sörbom, 1989), bifactor S-1 model analysis was implemented. Therefore, items which were in the PA factor were only loaded on the AF factor. In light of these findings, the bifactor S-1 model analysis was performed for both reverse-coded and non-reverse data sets. The results of the bifactor S-1 model analysis were given in Table 3.

Table 3

Bifactor S-1 model analysis for reverse coded and non-reverse coded data

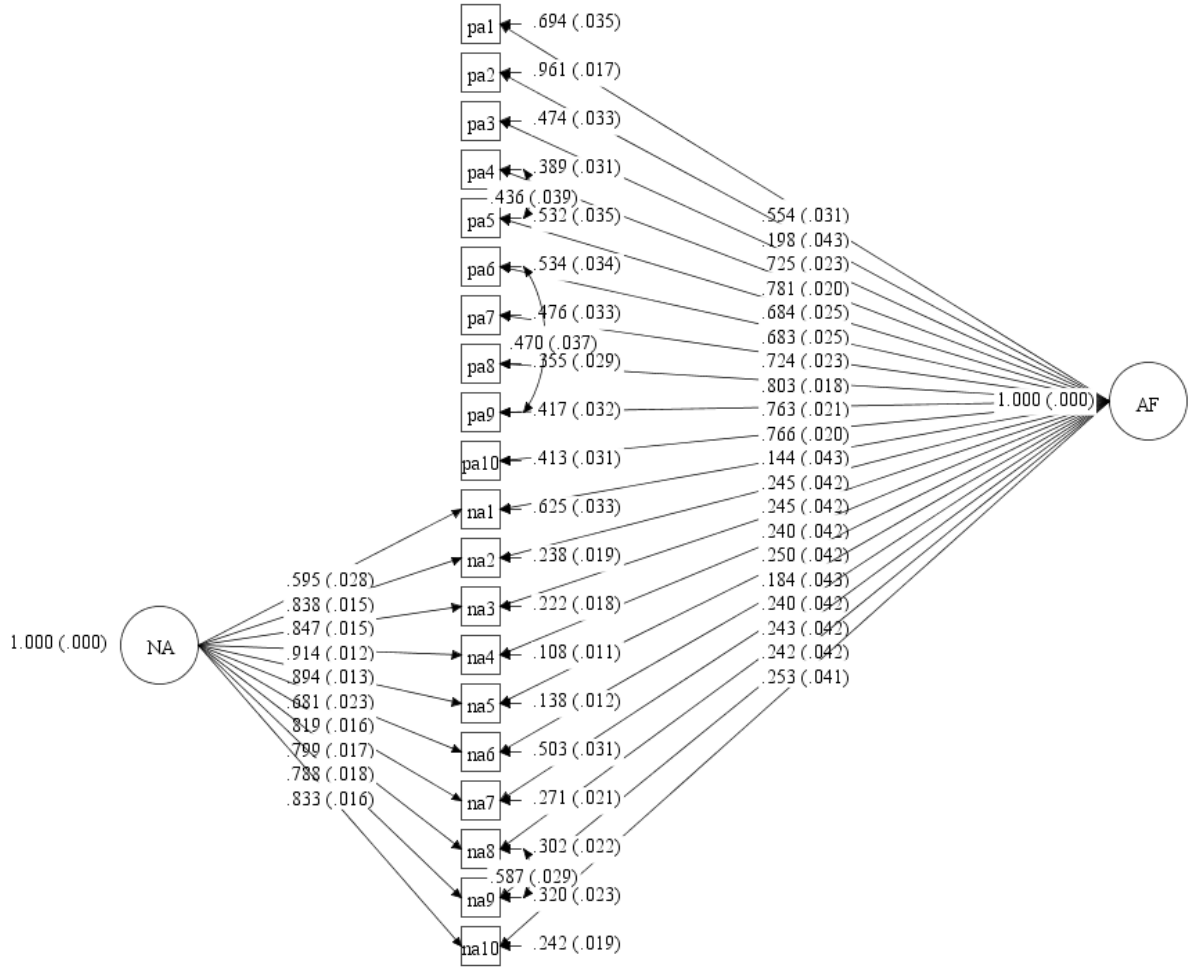
Coding type	Models	RMSEA	RMSEA CI		CFI	TLI	SRMR	Chi-Square
Non-reverse coded data	Without modification	0.111	0.106	0.116	0.880	0.857	0.102	1592.078 (160)
	With modification	0.087	0.082	0.092	0.927	0.912	0.100	1020.594 (156)
Reverse coded data	Without modification	0.113	0.107	0.118	0.877	0.854	0.097	9683.840 (190)
	With modification	0.089	0.084	0.095	0.924	0.908	0.095	876.259 (157)

When Table 3 was examined, it was seen that the bifactor S-1 model for reverse-coded data fitted with some modifications. As modifications, correlations of item PA4 with item PA5, item PA9 with item PA6, and item NA8 with item NA9 were allowed. Except for the RMSEA value, the model had an acceptable fit. Also, the chi-square values always had a problematic fit because of the sample size (Zimmer & Odum Institute, 2019). Therefore, at this point, for acceptable model-data fit, at least three model-data fit indices were considered. The big difference between the bifactor and bifactor S-1 models for reverse-coded data is that all path coefficients were positive in the bifactor S-1 model. This is an important finding for the compatibility of the bifactor S-1 model for the PANAS. The Bifactor s-1 model for PANAS was given in Figure 2.

In addition to the model-fit statistics, the factor loadings were examined. For the non-reverse coded data, items in the NA factor had negative factor loadings on the general factor (AF). This result shows that these items measure a feature opposite to the general factor (AF). However, in the bifactor S-1 analysis, items had higher loadings on their factors and increased unique variance. With the reverse-coded data, negatively loaded items in the NA factor were turned to positive. Moreover, items in the NA factor had lower factor loadings than 0.32 on the general factor (AF), which is in accordance with the studies on PANAS in the literature (Leue & Beauducel, 2011; Seib-Pfeifer et al., 2017). The path diagram in Figure 2 shows that there is not any negative factor loading. Only item PA2 had a lower factor loading among the PA items. This item also had a negative factor loading in existing studies in the literature (Leue & Beauducel, 2011; Seib-Pfeifer et al., 2017).

Figure 2

Bifactor S-1 model for reverse coded data



Conclusion, Discussion, And Recommendations

This study aims to ascertain whether or not the items in the NA factor should be reverse coded in the PANAS. Therefore, bifactor and bifactor S-1 model analyses were implemented. According to the results, for reverse-coded items, the bifactor S-1 model was seen as the better model for the PANAS. Additionally, in modeling unique variances of items with specific factors, the bifactor S-1 model performed well and also resolved the negative loading problem of the items on the general factor. Contrary to the bifactor S-1 model, the bifactor model had a poor fit for the PANAS. In accordance with the present result, in the studies using the bifactor model, better fit indices were obtained in most cases, while in a controversial study, it was seen that this model had a poor fit for the PANAS (for detailed information, see Flores-Kanter et al., 2021). Also, in the literature, studies which found a better fit with the bifactor model for the PANAS had a big problem, which was negative factor loadings (Leue & Beauducel, 2011; Seib-Pfeifer et al., 2017).

In this research, negative factor loadings were not obtained with the bifactor model, but the model fit could not be achieved. Therefore, in consideration of modification suggestions, the bifactor S-1 model analysis was made, the negative coefficients were eliminated, and the model fit was obtained. Besides, lower factor loadings were obtained with the bifactor model than with the bifactor s-1 model. The increase in factor loadings showed that the items in the specific factor had a remarkable contribution.

The lower factor loadings of the items in the specific factor reduced the importance of the specific factor.

A remarkable finding of this study is that the bifactor model revealed lower factor loadings on NA factor. According to Kula Kartal et al.'s (2022) research, the wording effect can cause lower factor loadings of negative items. In this study, lower factor loading may have arisen from the wording effect too.

Even though the items in the NA factor were reverse coded in the bifactor model analysis, the PA2 item belonging to the positive affect factor had a negative loading, as in other studies in the literature (Leue & Beauducel, 2011; Seib-Pfeifer et al., 2017). This may be because the expression "distressed" in the PA2 item might have been perceived negatively by the group. The adjective for this item should be reexamined in other studies.

One of the modification suggestions is to establish a correlation between the PA factor and the general factor. The need for reverse coding of negative items has emerged as a finding. The point to take into consideration with this finding, which should be noted by researchers who will bifactor model with PANAS, is that negative items should be reverse coded. In this research, modifications were implemented to obtain model-data fit. However, modified models need to estimate with an independent sample. In this way, the model can truly be termed "confirmed" technically. With new studies, the bifactor S-1 model for the PANAS should be reanalyzed with different samples.

In this study, negative items were reverse coded. But in the literature, there was a majority of views on reverse coding. Greenberger et al. (2003) recommend not to mix negative items with positive items because it creates a two-factor structure of the instrument based on the item wording difference (positively and negatively worded items), which is a threat to construct validity. Ibrahim (2001) also recommends not mixing negatively and positively worded items. Salazar (2015) states that although mixing can reduce the acquiescence bias, it causes a method effect, impairs factorial validity, and hurts internal consistency. Hartley (2013) also recommend that if researchers mix positively and negatively items, they should present results obtained from negatively worded items separately, instead of reverse-coding the data and combining them with the data obtained from positively worded items. Locker et al. (2007) recommend that when using a mixed format with the intent to reverse-code negatively worded items, make sure to use a symmetrical response scale with an equal number of anchors on the positive and negative sides of the scale.

Literature about reverse coding is debatable. Also, PANAS should not be used with negative factor loadings both general and specific factors. If researchers hesitate to use reverse coding, it is not recommended to use bifactor models with PANAS. Kula Kartal et al. (2022), in their studies recommended using the bifactor model to examine the wording effect of items. A further study should asses the wording effect of the NA factor on PANAS with the bifactor model.

Declaration

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

Bernard, H. R. (2013). *Social research methods: Qualitative and quantitative approaches*. Sage.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford.

Brown, J. D., & Marshall, M. A. (2001). Self-esteem and emotion: Some thoughts about feelings. *Personality and Social Psychology Bulletin, 27*(5), 575–584. https://doi.org/10.1177/0146167201275006

Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020). Application of the Bifactor S − 1 model to multisource ratings of ADHD/ODD symptoms: An appropriate bifactor model for symptom ratings. *Journal of Abnormal Child Psychology, 48*(7), 881-894. http://dx.doi.org/10.1007/s10802-019-00608-4

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactored tests: Implications for understanding multidimensional constructs and test interpretation. *Principles and Methods of Test Construction: Standards and Recent Advancements*. Hogrefe Publishers.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Erlbaum and Associates.

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. *Journal of Intelligence, 5*(3), 27. http://dx.doi.org/10.3390/jintelligence5030027

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment, 23*(3), 225-241. https://doi.org/10.1177/073428290502300303

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation, 14*(1), 20.

Ebesutani C., Smith A., Bernstein A., Chorpita B. F., Higa-McMillan C., & Nakamura B. (2011). A bifactor model of negative affectivity: Fear and distress components among younger and older youth. *Psychological Assessment, 23*(3), 679–691. http://dx.doi.org/10.1037/a0023234

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S − 1 model for individual clinical assessment. *Journal of Abnormal Child Psychology, 48*(7), 895-900. http://dx.doi.org/10.1007/s10802-020-00624-9

Flores-Kanter, P. E., Garrido, L. E., Moretti, L. S., & Medrano, L. A. (2021). A modern network approach to revisiting the Positive and Negative Affective Schedule (PANAS) construct validity. *Journal of Clinical Psychology, 77*(10), 2370-2404. http://dx.doi.org/10.1002/jclp.23191

Gaudreau, P., Sanchez, X., & Blondin, J.-P. (2006). Positive and negative affective states in a performance related setting. *European Journal of Psychological Assessment, 22*(4), 240–249. https://doi.org/10.1027/1015-5759.22.4.240

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research, 11*(1), 572-580. http://dx.doi.org/10.1086/208993

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences, 35*(6), 1241–1254. https://doi.org/10.1016/S0191-8869(02)00331-8

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2008). *Multivariate data analysis*. Prentice Hall Publisher.

Hartley, J. (2013). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology, 13*(1), 83–86. https://doi.org/10.1080/13645570802648077

Holzinger, K. J., & Swineford, F. (1937). The Bi-Factor method. *Psychometrika, 2*(1), 41-54.

Huebner, E. S., & Dew, T. (1995). Preliminary validation of the positive and negative affect schedule with adolescents. *Journal of Psychoeducational Assessment, 13*(3), 286–293. https://doi.org/10.1177/073428299501300307

Ibrahim, A. M. (2001). Differential responding to positive and negative items: The case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*, 497–500. https://doi.org/10.2466/pr0.2001.88.2.497

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7 user's reference guide*. Scientific Software.

Kula Kartal, S., Aybek, E. C., & Yaşar, M. (2022). Investigating the wording effect in scales based on different dimension reduction techniques. *Journal of Uludağ University Faculty of Education, 35*(1), 44-67. https://doi.org/10.19171/uefad.1033284

Killgore, W. D. S. (2000). Evidence for a third factor on the positive and negative affect schedule in a college student sample. *Perceptual and Motor Skills, 90*(1), 147–152. https://doi.org/10.2466/pms.2000.90.1.147

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Sage.

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

Lacobucci, D. (2010). Structural equations modelling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology, 20*(1), 90-98. https://doi.org/10.1016/j.jcps.2009.09.003

Leue, A., & Beauducel, A. (2011). The PANAS structure revisited: On the validity of a bifactor model in community and forensic samples. *Psychological Assessment, 23*(1), 215-225. http://dx.doi.org/10.1037/a0021400

Locker, D., Jokovic, A., & Allison, P. (2007). Direction of wording and responses to items in oral health-related quality of life questionnaires for children and their parents. *Community Dentistry and Oral Epidemiology, 35*(4), 255–262. https://doi.org/10.1111/j.1600-0528.2007.00320.x

Magyar-Moe, J. L. (2009). *Therapist's guide to positive psychological interventions*. Academic press.

Mihić, L., Novović, Z., Čolović, P., & Smederevac, S. (2014). Serbian adaptation of the Positive and Negative Affect Schedule (PANAS): Its facets and second-order structure. *Psihologija, 47*(4), 393–414. http://dx.doi.org/10.2298/PSI1404393M

Molwus, J. J., Erdogan, B., & Ogunlana, S. O. (2013). Sample size and model fit indices for structural equation modelling (SEM): The case of construction management research. In *ICCREM 2013: Construction and Operation in the Context of Sustainability* (pp. 338-347). http://dx.doi.org/10.1061/9780784413135.032

Ortuño-Sierra, J., Santarén-Rosell, M., de Albéniz, A. P., & Fonseca-Pedrero, E. (2015). Dimensional structure of the Spanish version of the positive and negative affect schedule (PANAS) in adolescents and young adults. *Psychological Assessment, 27*(3), e1–e9. https://doi.org/10.1037/pas0000107

Pires, P., Filgueiras, A., Ribas, R., & Santana, C. (2013). Positive and negative affect schedule: Psychometric properties for the Brazilian Portuguese version. *The Spanish Journal of Psychology, 16*, e58. https://doi.org/10.1017/sjp.2013.60

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data Yield Univocal Scale Scores. *Journal of Personality Assessment, 92*(6), 544-559. http://dx.doi.org/10.1080/00223891.2010.496477

Rush, J., & Hofer, S. M. (2014). Differences in within and between person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment, 26*(2), 462–473. https://doi.org/10.1037/a0035666.

Salazar, M. S. (2015). The dilemma of combining positive and negative items in scales. *Psicothema, 27*(2), 192–199. https://doi.org/10.7334/psicothema2014.266

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Seib-Pfeifer, L.-E., Pugnaghi, G., Beauducel, A., & Leue, A. (2017). On the replication of factor structures of the Positive and Negative Affect Schedule (PANAS). *Personality and Individual Differences, 107*(1), 201–207. https://doi.org/10.1016/j. paid.2016.11.053

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 183–206). Routledge/Taylor & Francis Group.

Stucky, B. D., Edelen, M. O., Vaughan, C. A., Tucker, J. S., & Butler, J. (2014). The psychometric development and initial validation of the DCI-A short form for adolescent therapeutic community treatment process. *Journal of Substance Abuse Treatment, 46*(4), 516-521. https://doi.org/10.1016/j.jsat.2013.12.005

Vera-Villarroel, P., Urzúa, A., Jaime, D., Contreras, D., Zych, I., Celis-Atenas, K., Silva, J. R., & Lillo, S. (2017). Positive and Negative Affect Schedule (PANAS): Psychometric properties and discriminative capacity in several chilean samples. *Evaluation & the Health Professions, 42*(4), 473–497. https://doi.org/10.1177/0163278717745344

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070. https://doi.org/10.1037/0022- 3514.54.6.1063

Zampetakis L. A., Lerakis M., Kafetsios, K., & Moustakis, V. (2015). Using item response theory to investigate the structure of anticipated affect: Do self-reports about future affective reactions conform to typical or maximal models? *Frontiers in Psychology, 6*, 1438. https://doi.org/10.3389/fpsyg.2015.01438

Zimmer, C., & Odum Institute (2019). Learn to perform confirmatory factor analysis in Stata with data from general social survey (2016). In SAGE research Methods Datasets Part 2. SAGE Publications. https://dx.doi.org/10.4135/9781529700091

Appendix

Appendix A

Table A1

Bifactor path coefficients for not reverse coded data

	Without modification			With modification		
Items	AF	PA	NA	AF	PA	NA
PA1	0.603	0.523		0.601	0.522	
PA2	0.638	0.247*		0.658	0.290*	
PA3	0.096	0.714		0.112*	0.747	
PA4	0.346	0.746		0.369	0.788	
PA5	0.261*	0.706		0.280*	0.744	
PA6	0.335	0.709		0.286*	0.651	
PA7	0.337	0.706		0.339	0.717	
PA8	0.277*	0.814		0.236*	0.768	
PA9	0.303	0.775		0.238*	0.707	
PA10	0.252*			0.240*	0.739	
NA1	0.699		0.143*	0.705		0.117*
NA2	0.822		0.373	0.839		0.337
NA3	0.790		0.491	0.814		0.455
NA4	0.833		0.491	0.856		0.451
NA5	0.772		0.563	0.800		0.524
NA6	0.820		0.162*	0.827		0.129*
NA7	0.810		0.374	0.828		0.339
NA8	0.948		0.140*	0.953		0.102*
NA9	0.939		0.140*	0.945		0.100*
NA10	0.846		0.326	0.861		0.289*

*<0.320

Table A2

Bifactor path coefficients for reverse coded data with modifications

	Without modification			With modification		
Items	AF	PA	NA	AF	PA	NA
PA1	0.699	0.401		0.699	0.408	
PA2	-0.383	0.358		-0.380	0.360	
PA3	0.559	0.596		0.561	0.592	
PA4	0.444	0.731		0.444	0.716	
PA5	0.444	0.651		0.439	0.621	
PA6	0.437	0.642		0.431	0.614	
PA7	0.390	0.674		0.398	0.684	
PA8	0.495	0.690		0.499	0.693	
PA9	0.500	0.669		0.499	0.654	
PA10	0.477	0.662		0.483	0.668	
NA1	0.687		0.172*	0.687		0.169*
NA2	0.915		0.005*	0.915		0.001*
NA3	0.925		-0.027*	0.925		-0.030*
NA4	0.967		-0.002*	0.967		-0.007*
NA5	0.961		-0.035*	0.961		-0.040*
NA6	0.771		0.224*	0.772		0.221*
NA7	0.901		0.039*	0.901		0.036*
NA8	0.876		0.387	0.878		0.383
NA9	0.869		0.380	0.871		0.376
NA10	0.906		0.169*	0.907		0.165*

*<0.320

Table A3

Bifactor S-1 path coefficients for non-reverse coded data

	Without modification		With modification	
Items	AF	NA	AF	NA
PA1	0.537		0.546	
PA2	0.141*		0.159*	
PA3	0.686		0.709	
PA4	0.748		0.768	
PA5	0.689		0.691	
PA6	0.712		0.662	
PA7	0.701		0.723	
PA8	0.825		0.783	
PA9	0.787		0.720	
PA10	0.737		0.748	
NA1	-0.108*	0.593	-0.113*	0.602
NA2	-0.218*	0.836	-0.223*	0.836
NA3	-0.221*	0.867	-0.220*	0.843
NA4	-0.265*	0.904	-0.270*	0.909
NA5	-0.243*	0.875	-0.248*	0.888
NA6	-0.159*	0.717	-0.169*	0.693
NA7	-0.197*	0.829	-0.203*	0.820
NA8	-0.233*	0.835	-0.239*	0.822
NA9	-0.255*	0.817	-0.262*	0.813
NA10	-0.243*	0.830	-0.251*	0.841

*<0.320

Table A4

Bifactor S-1 path coefficients for reverse coded data

	Without modification		With modification	
Items	AF	NA	AF	NA
PA1	0.543		0.554	
PA2	0.203*		0.198	
PA3	0.727		0.725	
PA4	0.799		0.781	
PA5	0.717		0.684	
PA6	0.717		0.683	
PA7	0.714		0.724	
PA8	0.798		0.803	
PA9	0.780		0.763	
PA10	0.756		0.766	
NA1	0.141*	0.602	0.144*	0.595
NA2	0.238*	0.836	0.245*	0.838
NA3	0.242*	0.843	0.245*	0.847
NA4	0.234*	0.909	0.240*	0.914
NA5	0.245*	0.888	0.250*	0.894
NA6	0.172*	0.693	0.184*	0.681
NA7	0.234*	0.820	0.240*	0.819
NA8	0.237*	0.822	0.243*	0.799
NA9	0.236*	0.813	0.242*	0.788
NA10	0.250*	0.841	0.253*	0.833

* <0.320

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 256-268

The Effect of Aberrant Responses on Ability Estimation in Computer Adaptive Tests

Sebahat GÖREN* Hakan KARA** Başak ERDEM KARA***
Hülya KELECİOĞLU****

Abstract

In computer adaptive test (CAT), aberrant responses caused by some factors such as lucky guesses and carelessness errors may cause significant bias in ability estimation. Correct responses resulting from lucky guesses and false responses resulting from carelessness or anxiety may reveal aberrant responses and the impact of these types of aberrant responses may cause an erroneous estimation of the examinee's actual ability because they do not reflect the examinee's actual knowledge. In this study, the performances of regarding ability estimation were examined comparatively in the context of CAT simulations in case of aberrant responses.Under different conditions, twelve different CAT simulations were conducted with 10 replications for each of the conditions. Correlation, RMSE, bias, and mean absolute error (MAE) values were calculated and interpreted for each condition. Results generally indicated that the 4PL IRT model provided a more efficient and robust ability estimation than the 3PL IRT model and the 4PL model increased the precision and effectiveness of the CAT applications.

Keywords: Computer adaptive tests (CAT), 3PL IRT model, 4PL IRT model, aberrant responses, early mistake

Introduction

Nowadays, many achievement tests, most of which are applied as multiple-choice, are carried out for different purposes such as selection, placement, classification, and evaluation, especially in the field of education. The process of preparing, applying, and evaluating these tests in order to estimate the latent characteristic of individuals in the most appropriate way also changes with the advancement of knowledge and technology. In recent years, the popularity and use of CAT applications, which minimize random errors by providing items appropriate to the individual's ability level, and thus provide the opportunity to reach more accurate information, has increased. So that, individuals only answer items that are appropriate for their ability levels, the length of the test is shortened and the test duration decreases (Thompson, 2009; Wainer, 2000; Weiss, 2004). CAT applications are mostly based on IRT models and IRT models allow for estimation of abilities and comparisons between individuals, even when individuals answer different items at different difficulty levels The test algorithm applied in CAT consists of three basic steps; the starting rule, the progression rule, and the termination rule (Wainer, 2000). Those steps include the application of a predetermined starting rule using an item pool of a sufficient qualified and number of items (starting rule), the selection of the most appropriate item for individual's ability level from the pool based on the temporary ability level calculated after each answered item (progression rule) and the termination of the test based on a specified termination rule (Segall, 2004; Thompson & Weiss, 2011).

Individuals' answers on multiple choice tests are classified into three categories; responses reflecting true ability, correct responses given by chance (lucky guesses), and false responses resulting from anxiety, carelessness, or distraction (Liao et al., 2012). The last two categories, which contain unusual

* Research Assistant, Hacettepe University, Faculty of Education, Ankara-Turkey, sebahatgoren@gmail.com, ORCID ID: 0000-0002-6453-3258
** PhD student., Hacettepe University, Faculty of Education, Ankara-Turkey, hakankaraodtu@gmail.com, ORCID ID: 0000-0002-2396-3462
*** Assist. Prof. Dr., Anadolu University, Faculty of Education, Eskişehir-Turkey, basakerdem@anadolu.edu.tr, ORCID ID: 0000-0003-3066-2892
**** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Gören, S., Kara, H., Erdem-Kara, B., & Kelecioğlu, H. (2022). The effect of aberrant responses on ability estimation in computer adaptive tests. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 256-268. https://doi.org/10.21031/epod.1067307

Received: 02.02.2022
Accepted: 10.07.2022

answers that we are not used to, do not reflect the true ability level of the individual and cause an erroneous estimate of the individual's true ability. The effect of these abnormal responses is limited due to equal weighting of items in traditional tests based on classical test theory. However, IRT is highly sensitive to that kind of response disturbance since it is a statistical method based on an examinee's response to explain the ability level (Magis, 2014). The existence of aberrant responses may cause a strongly biased estimation of true underlying ability and may jeopardize the accuracy of measurements and invalidate the IRT use (Jia et al., 2019). Since CAT applications are also based on IRT models, they are also open to the same kind of biased estimations and erroneous measurements. Rulison and Loken (2009) stated that chance (luck) factors and attention errors that occur especially at the beginning of a test in CATs may have a significant effect on test results. To solve this problem, Barton and Lord (1981) developed 4PL IRT model from IRT models by adding the inattention (carelessness) parameter (d_i) to 3PL IRT model.

With 4PL IRT model, the probability of individuals with high ability levels giving wrong answers for easy items because of the factors such as carelessness, fatigue, or anxiety is calculated. Inattention parameter (d_j) allows the upper asymptote to get values smaller than 1.00 and differ between 0.00 and 1.00 theoretically. With the inclusion of an upper asymptote with a value less than 1.00, it is allowed that the place of a high-ability individual does not change significantly in the ability scale in case of a false response to an easy item. Barton and Lord (1981) conducted analysis by fixing upper asymptote values 1.00, 0.99, and 0.98 respectively. In other words, they specified a common upper asymptote value for all items and did not mention the freely estimating d parameter (Waller & Reise, 2010).

With CAT applications, making the temporary ability estimations of each individual after each item and selection of item based on those temporary ability levels makes the use of 4PLM quite meaningful. Since 4PLM aims to estimate the ability of individuals with the high ability with the least error, it is more affected by errors such as carelessness, especially at the beginning of the test, which causes estimation bias (Rulison & Loken, 2009). For example, when the d parameter of the item is in the range of 0-1, when the individual answers this item incorrectly, there will be a decrease in the individual's ability level and this decrease will be less than the other incorrectly answered items. Thus, an item more appropriate for the ability level of the individual can be selected from the item pool later.

Rulison and Loken (2009) made ability estimations with ordinary (not intervened) performance, intervened performances such that individuals intentionally answered the first two items incorrectly and again intentionally answered the first two items correctly in order to investigate the effect of upper asymptote on ability estimation in CAT applications under 3PLM and 4PLM by conducting a simulation study. It was concluded that 4PLM may reduce the estimation error for high-ability examinees who answered the first two items incorrectly. In that study, the analyzes were made by fixing the d parameter to 0.98. Besides, Loken and Rulison (2010) indicated how to make parameter estimates with 4PLM by using both real and simulation data, and at the same time, they compared models by making estimations with 2PLM and 3PLM. While the correlation coefficients were similar for these two data types, the error values were at the lowest level for 4PLM. When estimated item parameters under 4PLM were examined it was observed that d parameter got values between 0.72 and 0.89. On the other hand, Waller and Reise (2010), used Low Self-Esteem Scale of the Minnesota Multiphasic Personality Inventory (MMPI-A) to examine compatibility of data with 4PLM. In that study, ability estimations were made under 3PLM and 4PL models and it resulted that ability estimation did not differ significantly, but only when the relationship between estimations of individuals at high ability level (with low self-esteem) was considered, model selection led to a difference. When the standard errors related to estimations were examined, it was observed that 4PLM had a more accurate estimation with less error. In addition, Liao et al. (2012) conducted a simulation study to compare the measurement precision and efficiency of 3PL and 4PL models under ordinary and poor-start testing conditions. CAT application was carried out under two different conditions; ordinary (not intervened) and the condition in which the first two items were intentionally evaluated as incorrect. Besides, estimations were made according to both 3PL and 4PL models in each of these conditions, and results were compared. When estimation was made with 4PLM, d parameter was fixed to di = 0.98 for all items. It was found that there was no significant difference between models in both ordinary and intervened conditions. When ability estimations were made under 4PL model, the error level was significantly lower than the error on 3PL model.

When individuals take achievement tests in CAT, they can give incorrect answers to items because of some factors such as anxious, careless, and poor testing conditions although they know the correct answer. Those answers are some examples of aberrant responses, and those kinds of answers may cause significant bias in ability estimation especially when they are done at the beginning of the test since estimated thetas do not reflect the examinee's actual knowledge. So, the impact of these types of aberrant responses may cause an erroneous estimation of the examinee's actual ability because they do not reflect the examinee's actual knowledge. To cope with the effect of that aberrant responses, the use of 4PL IRT model is suggested in the literature. It is stated that 4PL IRT model may provide a more efficient and robust ability estimation than the 3PL IRT model in the CAT applications. Rulison and Luken (2009) stated that aberrant responses may influence the CAT results especially if they occur as early mistakes. 4PL IRT model improves the efficiency and precision of CAT under both ordinary conditions and the existence of aberrant responses. It helps to reduce the precision and efficiency degradation caused by careless mistakes (Liao et al., 2012). Despite of that, studies regarding the 4PL IRT model used for aberrant responses on computer adaptive tests (CAT) are limited to a few numbers of studies in the literature (Liao et al., 2012; Rulison & Loken, 2009). In the context of this study, the performances of 3PL and 4PL IRT models on ability estimation were investigated under the conditions in which ordinary (not intervened), the first item was intentionally evaluated as incorrect, and the first two items were intentionally evaluated as incorrect according to different termination rules. The results of the study are likely to contribute to the literature focusing on aberrant responses in computer adaptive test applications.

Purpose of the Study

In the context of this study, the performance of 3PLM and 4PLM as an error correction mechanism on CAT in case of the existence of aberrant responses was investigated under 12 conditions. In the study conducted for this purpose, an answer to the following research question was sought:

How do the performances of 3PL and 4PL models regarding ability estimation in CAT applications differ according to different response behaviors (ordinary, incorrect answers to the first item and the first two items) and different termination rules (fixed length and varying length)?

Method

In this study, performances of 3PL and 4PL models regarding ability estimations in CAT applications were examined under different response behaviors and different termination rules generated simulatively.

Data Generation

In this study, 13 ability levels ranging from -3.0 to +3.0 in equally spaced intervals of .5 were specified and the ability parameters of a total of 2600 individuals, including 200 individuals for each of the 13 levels were generated two different item pools each consisting of 300 items were generated by using 3PL and 4PL IRT models. The a, b, and c parameters of items in the item pool were generated by using log-normal distribution L(0, .25), uniform distribution U[-3, +3], and uniform distribution U[0, .25] respectively. The d parameter in the pool generated using 4PL IRT model was taken as 0.98, based on the findings obtained by Rulison and Loken (2009). After the generation of ability parameters and construction of the item pool, response patterns of examinees were generated and continued with CAT simulation.

Procedure

In CAT simulation, two different IRT models (3PL and 4PL); two different termination rules (30 items and SE< .30); and three different response behaviors (ordinary, poor-start1, and poor-start2) were used and 12 conditions (***2 IRT models x 3 response behaviors x 2 termination rules = 12 conditions***) were investigated in total. Twelve different CAT applications were applied to each participant for each of the conditions with 10 replication. Individuals' response to the first item in tests with Poor-start1 response behavior and to the first two items in tests with Poor-Start2 response behavior was intentionally evaluated as "incorrect" regardless of their true ability level. Those manipulated items were of medium difficulty. The remaining items were answered according to their true (actual) ability levels. Examinees answered all items according to their true ability levels in ordinary response behavior test conditions. The 12 different conditions are given below:

- O3CAT: Fixed-length CAT based on the 3PL model under ordinary response behavior
- O4CAT: Fixed-length CAT based on the 4PL model under ordinary response behavior
- $O3CAT_s$: Variable-length CAT based on 3PL model under ordinary response behavior
- $O4CAT_s$: Variable-length CAT based on 4PL model under ordinary response behavior
- P1_3CAT : Fixed-length CAT based on 3PL model under Poor-Start1 response behavior
- P1_4CAT : Fixed-length CAT based on 4PL model under Poor-Start1 response behavior
- $P1\_3CAT_s$: Variable-length CAT based on 3PL model under Poor-Start1 response behavior
- $P1\_4CAT_s$: Variable-length CAT based on 4PL model under Poor-Start1 response behavior
- P2_3CAT: Fixed-length CAT based on 3PL model under Poor-Start2 response behavior
- P2_4CAT: Fixed-length CAT based on 4PL model under Poor-Start2 response behavior
- $P2\_3CAT_s$: Variable-length CAT based on 3PL model under Poor-Start2 response behavior
- $P2\_4CAT_s$: Variable-length CAT based on 4PL model under Poor-Start2 response behavior

For each condition, the ability level for starting rule was specified as '0' and Maximum Fisher Information (MFI) method was used as the item selection method. In order to prevent the same item taking by each individual, randomesque method was used with a 5-item group. The expected a posteriori (EAP) method was preferred for ability estimation and a 0.50 value was used for item exposure.

Data Analysis

In data analysis, performances of 3PL and 4PL models regarding ability estimations were compared with the help of examining values regarding measurement precision. Those values which were the Pearson correlation coefficient between true ability and estimated ability levels, bias, and RMSE were calculated by taking the average of 10 replications for each condition.

RMSE was calculated by using the following formula. It is the root mean square error on all conditions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)^2}{n}} \quad (1)$$

Bias is the mean difference between an individual's true ability and estimated ability level as a result of simulation (Miller & Miller, 2004). It is calculated by using the following formula;

$$Bias = \frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)}{n} \quad (2)$$

On the other hand, mean absolute error (MAE) is the mean average difference between individuals' estimated ability level and true ability level.

$$MAE = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_i|}{n} \quad (3)$$

Results

In the context of this study, performances of 3PL and 4PL models regarding ability estimations were examined under different response behaviors (Ordinary, Poor-Start1, and Poor-Start2) and different termination rules (30 items and SE< .30). Correlation between true and estimated ability levels, RMSE, bias and mean absolute error (MAE) values were calculated and presented in Table 1. Besides, values in Table 1 were visualized and graphs regarding correlation, RMSE, and MAE for each test were presented in Figure 1. Obtained results were interpreted by considering both the table values and figures together.

Table 1

Correlation (r), RMSE, Bias and MAE Values of Tests

Response Behaviors	Test	r	RMSE	Bias	MAE
Ordinary	O3CAT	0.984	0.358	-0.007	0.285
	O4CAT	0.983	0.369	-0.01	0.293
	$O3CAT_S$	0.985	0.338	0.002	0.269
	$O4CAT_S$	0.986	0.335	0	0.266
Poor - Start1	P1_3CAT	0.982	0.419	-0.11	0.335
	P1_4CAT	0.981	0.419	-0.086	0.333
	$P1\_3CAT_S$	0.985	0.412	-0.14	0.333
	$P1\_4CAT_S$	0.985	0.375	-0.104	0.302
Poor - Start2	P2_3CAT	0.982	0.435	-0.132	0.347
	P2_4CAT	0.981	0.445	-0.128	0.352
	$P2\_3CAT_S$	0.985	0.378	-0.099	0.302
	$P2\_4CAT_S$	0.985	0.377	-0.102	0.299

When Table 1 and Figure 1a were examined, it was observed that the correlation between individuals' true ability levels generated before simulation and estimated ability levels because of CAT simulations is similar and high (≥.981) across tests. However, the correlation value was higher for variable-length tests compared to fixed-length tests regardless of response behavior and IRT model. Besides, the 3PL model has a higher correlation value for all answering behaviors although it has quite close values with the 4PL model in fixed-length tests. In the variable-length tests, the 4PL model has a higher correlation in ordinary response behavior, while it is almost equal in other conditions.

When Table 1 and Figure 1b were examined, 4PL is the model with a lower error value for all response behaviors in the variable-length test. In the fixed-length tests, the RMSE values for Poor-Start1 condition were equal for both models while the 3PL model had a lower error value in other conditions.

Bias values in Table 1 revealed that the 4PL model had lower bias values than 3PLM for all conditions regardless of response behaviors and termination rules. Except for ordinary response behavior, negative bias values were obtained in the other 10 conditions. Based on this, it can be said that the true ability of

individuals was underestimated in these tests, but this effect is lower for the 4PL model. In order to calculate and interpret the degree of estimation bias, MAE values were calculated and interpreted.

When Table 1 and Figure 1c were examined together, it was seen that the lowest MAE values were obtained in the 4PL model in all response behaviors for the variable-length test. In the fixed-length test, the 3PL model had a lower MAE value in the ordinary and PoorStart-2 response behaviors, and the 4PL model in the other condition.

Based on all these findings, it can be interpreted that 4PLM offers higher measurement precision for variable-length tests on all response behaviors in general. However, 3PLM gives better results for fixed-length tests. For more detailed interpretations, changes in those values on different ability levels were examined and obtained results were presented in Figure 2 - Figure 7.

As seen in Figure 2, under ordinary response behavior, in fixed length tests RMSE values of O3CAT were generally lower across ability levels. That is, in case of using the 3PLM, obtained values were lower compared to the O4CAT. For variable length tests (O3CATs - O4CATs) 3PL and 4PL models presented similar results. Based on these graphs, it is not appropriate to interpret that one model is superior to another. When graphs were examined in general, it can be shown that RMSE values were lower for variable-length tests compared to fixed-length tests across ability levels. Besides, while the RMSE values of the tests at low ability levels were closer to each other, as the skill level increased, the RMSE values also moved away from each other.

Figure 1

Correlation, RMSE and MAE Values of Tests

Figure 1a. Correlation

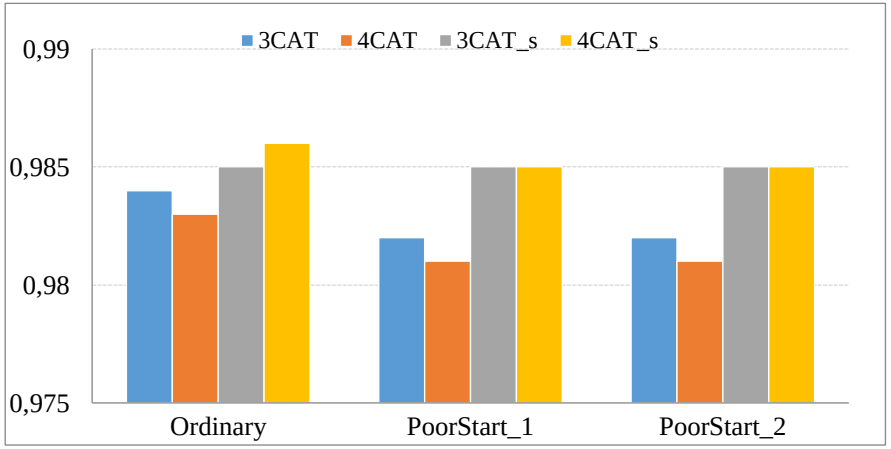


Figure 1b. RMSE

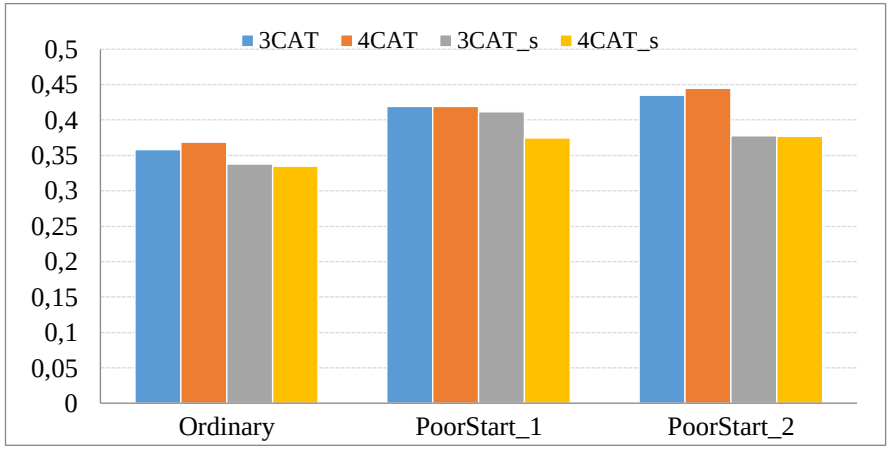


Figure 1c. Mean Absolute Error (MAE)

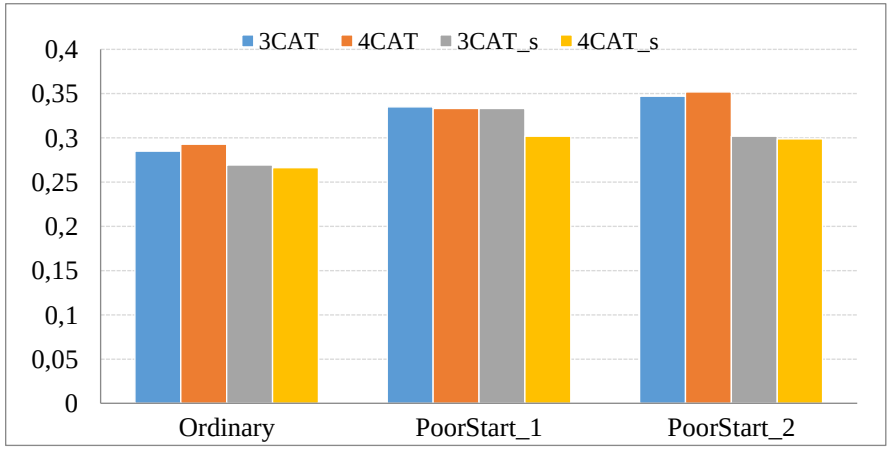
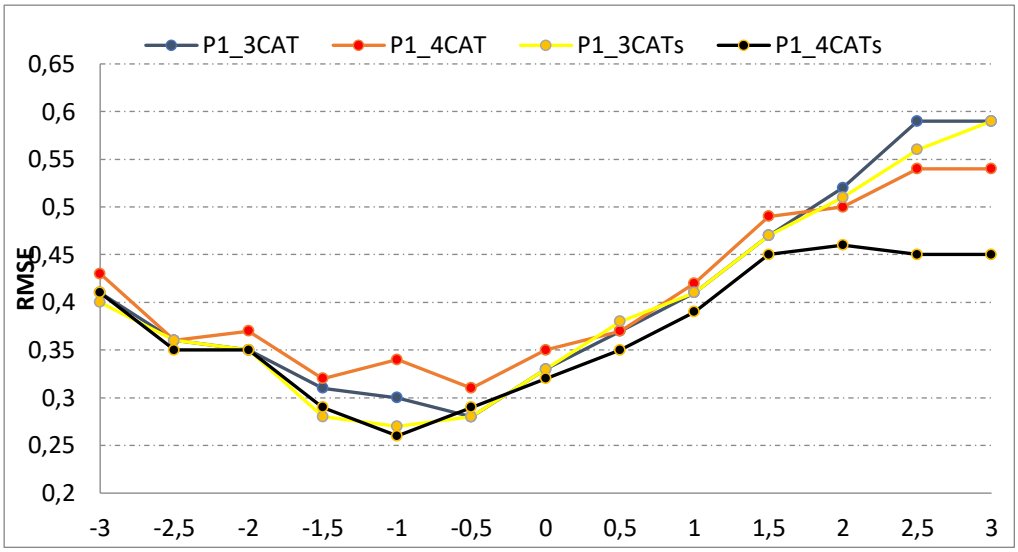


Figure 3

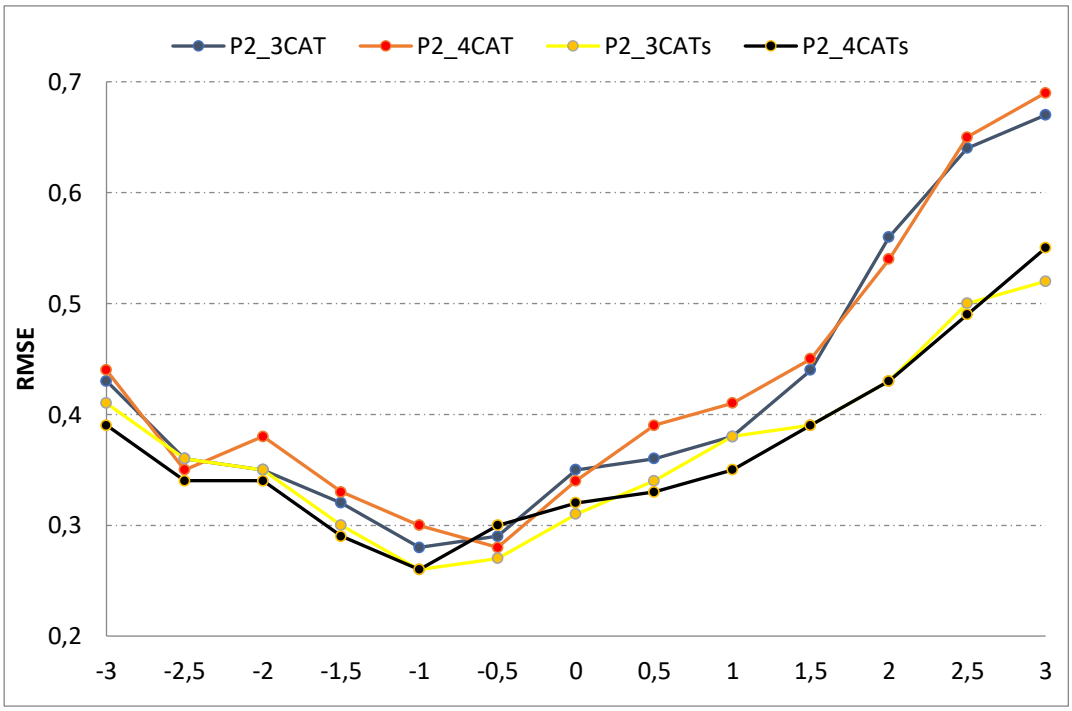
RMSE Values Across Ability Levels (Poor-Start1)



As seen in Figure 3, when RMSE values of fixed-length tests (P1_3CAT - P1_4CAT) were compared under Poor-Start1 response behavior; the values of P1_3CAT at low and medium ability levels and the values of P1_4CAT at high ability levels were lower. That is, the ability of individuals with high ability was estimated more accurately with 4PLM. For variable-length tests (P1_3CATs - P1_4CATs), RMSE values of P1_4CATs were lower at high ability levels compared to P1_3CATs and they have similar values at low and medium ability levels. Based on that graph, under Poor-Start1 response behavior, it can be interpreted that 4PLM gave better results. In addition, it was observed that RMSE values of those tests were closer at low and medium ability levels and getting farther at high ability levels.

Figure 4

RMSE Values Across Ability Levels (Poor-Start2)

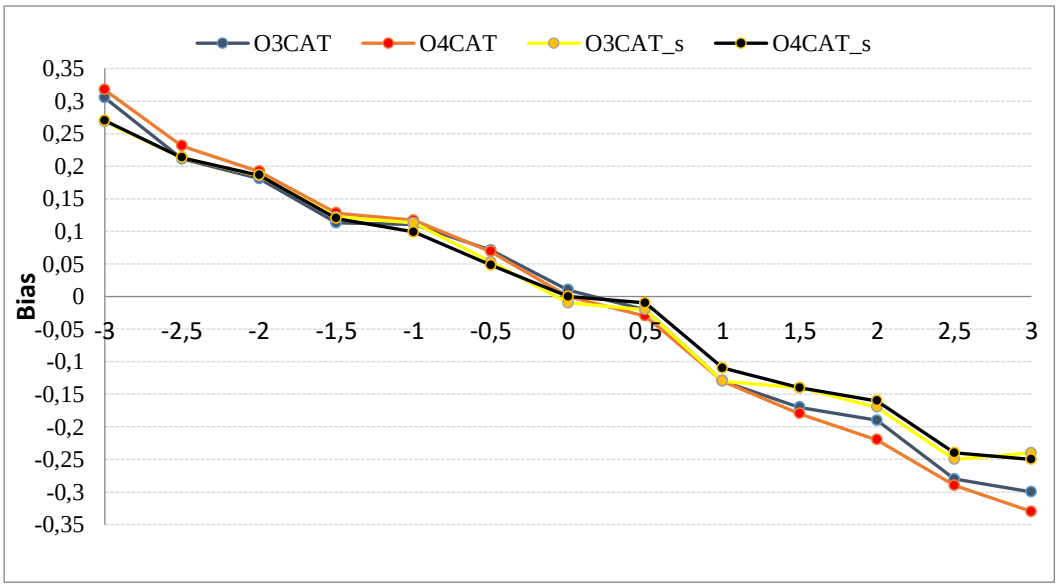


As seen in Figure 4, in terms of RMSE values, both fixed-length tests (P2_3CAT - P2_4CAT) and variable-length tests (P2_3CATs - P2_4CATs) had similar results across ability levels under Poor-Start 2 response behavior. Therefore, it can be said that models do not give superiority to each other. When the graph was examined in general, it can be observed that smaller RMSE values were obtained for variable-length tests compared to fixed-length tests at high ability levels. However, similar RMSE values were found at low and medium ability levels.

As seen in Figure 5, under ordinary response behavior, bias values of tests (O3CAT, O4CAT, O3CATs, and O4CATs) were similar at low and medium ability levels. However, at high ability levels and the lowest ability level ($\theta = -3$), fixed-length tests (N3CAT and N4CAT) were similar to each other, and variable-length tests (O3CATs and O4CATs) were similar to each other in terms of bias values. Therefore, under ordinary response behavior, 3PL and 4PL models were not superior to each other in terms of bias. When the graph was examined in general, the ability of individuals at low ability levels was overestimated (estimated higher than it really is) and the ability of individuals with high ability levels was underestimated (estimated lower than it really is). Besides, at high ability levels, lower bias values were estimated at variable-length tests compared to fixed-length tests, and the ability of individuals was estimated more accurately.

Figure 5

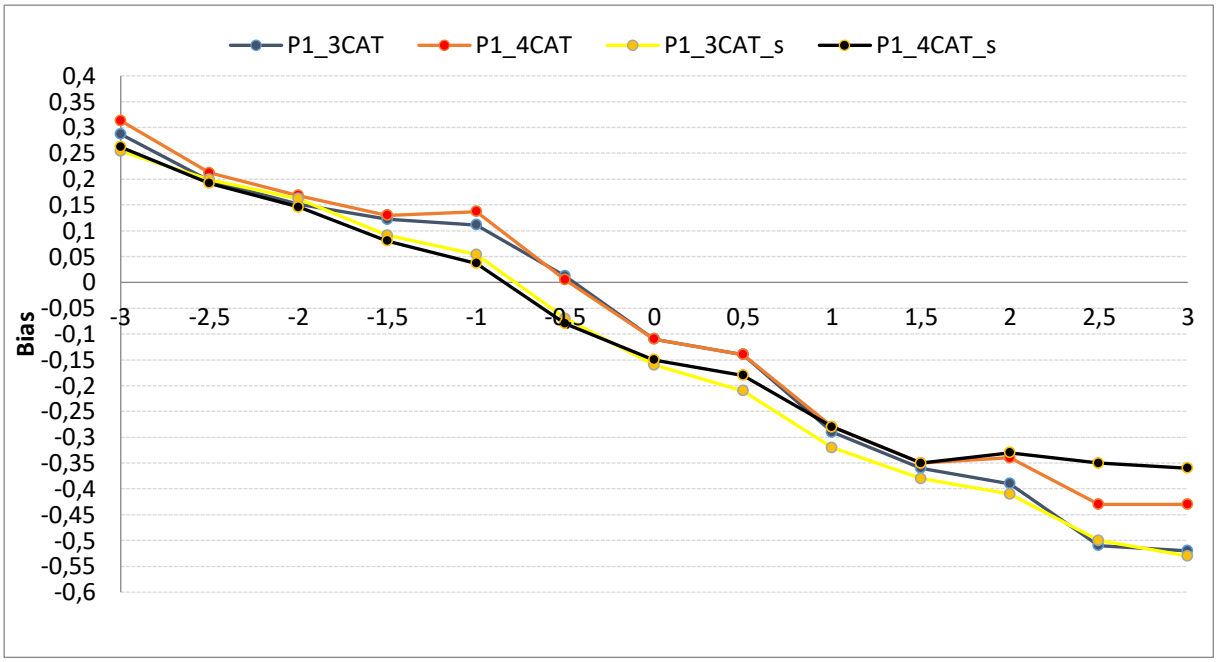
Bias Values Across Ability Levels (Ordinary)



According to Figure 6, when the bias values of fixed-length tests (P1_3CAT - P1_4CAT) were compared under the Poor-Start1 response behavior; P1_4CAT values were found to be smaller at higher skill levels, while similar values were calculated at other skill levels. Similarly, variable-length tests (P1_3CATs - P1_4CATs) had lower estimation bias at high ability levels compared to P1_3CAT. Therefore, it can be interpreted that 4PL was better at estimating high-level individuals' ability under Poor-Start1 response behavior.

Figure 6

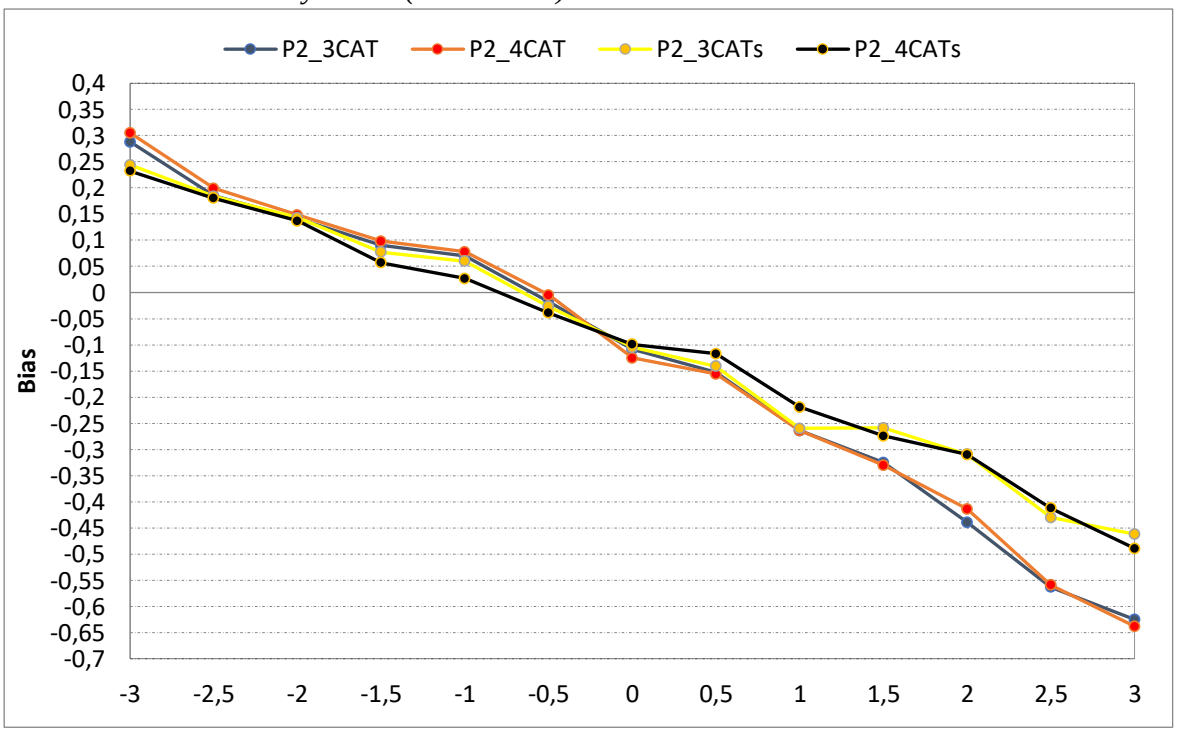
Bias Values Across Ability Levels (Poor-Start1)



When Figure 7 was examined, fixed-length tests (P2_3CAT - P2_4CAT) were similar to each other, and variable-length tests (P2_3CATs - P2_4CATs) were similar to each other in terms of bias values at high ability levels under Poor-Start2 response behavior. It was seen that the estimation bias of tests was closer at other ability levels. Accordingly, the models could not provide superiority to each other in the ability estimation of individuals with Poor-Start2 response behavior. In addition, lower bias values were generally estimated in variable-length tests compared to fixed-length tests, and individuals' abilities were estimated more accurately.

Figure 7

Bias Values Across Ability Levels (Poor-Start2)



Discussion and Conclusion

In the context of this study, 12 different CAT simulations with different response behaviors (ordinary-not intervened, the intentionally incorrect answer to the first item, and intentionally incorrect answer to the first two items) and different termination rules (30 items and SE<.30) in order to interpret the performances of 3PL and 4PL models regarding ability estimation.

Regardless of response behavior and IRT model, correlation values between individuals' true abilities (generated before simulation) and estimated abilities were higher and RMSE and bias were lower at variable-length tests compared to fixed-length tests. The clear conclusion of this study is that variable-length test performed generally better than fixed-length test in terms of correlation, bias, RMSE, and MAE through all conditions. That finding is consistent with the results of Babcock and Weiss (2012) stating that variable-length CATs performed either slightly better or comparable to the fixed-length test. The reason for the improved efficiency of variable-length tests may be that they allow using of more items which may yield better θ estimates.

In addition, it was concluded that the usage of 4PLM at variable-length tests provided higher measurement precision but 3PLM worked better at fixed-length tests. For the variable-length tests, that is an expected result since it is known that the 4PL model provides for more robust estimations in case of violations of IRT's well-known assumptions (Ackerman, 1989). As a result, aberrant errors that are inconsistent with an examinee's ability affect the ability estimate in 4PL model less than in the 3PL model. Similar to that result, Liao et al. (2012) used variable-length tests (SE<.30) in order to compare models with CAT applications and they concluded that 4PLM gave better results. On the other hand, it is not clear why 3PL model provided higher measurement efficiency for the fixed-length test. Although a certain explanation cannot be made for that, the possible reason may be the test length. Variable-length tests required more items than the fixed ones through all conditions. 4PLM model may have required to use more items to work better.

In addition to those findings, our results also demonstrated that the difference in RMSE and bias between fixed-length and variable-length tests were moving away from each other especially at higher ability levels for each response behavior. This result indicated that the measurement efficiency of fixed and variable-length tests was becoming distant toward high ability levels. In most of the conditions, variable-length tests presented lower RMSE and bias values than fixed-length tests which indicated that variable-length tests had better measurement efficiency for high-ability levels not only in the case of aberrant responses but also the ordinary condition. In addition to that, 4PLM presented either too similar or better results at high ability levels for variable-length tests. The ability of high-level individuals was estimated more accurately with 4PLM and that result is in concordance with the findings in the literature (Liao et al.,2012; Loken & Rulison, 2010; Rulison & Loken, 2009; Waller & Reise, 2010). The reason for the improved efficiency that 4PLM provided for high ability may be that 4PLM is more robust than the 3PLM and the upper asymptote value of 1 in 3PLM failed to accommodate the aberrant responses of high-ability students. On the other hand, the same comments may not be applicable for the fixed-length test. Under ordinary response behavior, 3PLM presented better results compared to 4PLM at estimating the ability of high-ability individuals. Under Poor-Start1 response behavior, it was observed that 4PLM worked better at high level individual's ability estimation. Under Poor-Start2 behavior, it is difficult to say that one model is superior to the other since they behaved in a very similar way. The obtained results are not sufficient to make any explanation or comment for fixed-length tests.

As a result, factors such as being anxious, being affected by poor testing conditions, lack of computer familiarity or misreading the question may cause individuals to give wrong answers to the questions that should have been answered correctly. If those mistakes are made at the beginning of the test, the ability of the individuals can be underestimated under the 3PL model. The results of this study have shown that the 4PL model increases the effectiveness and precision of the CAT in case of aberrant responses, especially for variable-length tests. A more detailed investigation is needed for the fixed-length tests. Additional research should be overtaken to examine this issue under different conditions. Another suggestion for further research is that aberrant responses were handled only as early mistakes in the context of that study. Further research can be made to investigate the effect of aberrant responses which are not an early mistake.

This research does have limitations that could limit the generalizability of our results. The item pools and examinees' ability parameters were specified by the researchers in accordance with the literature and replications were conducted but it is possible to get different results with different item banks and ability distributions.

Declarations

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

Ethical Approval: Data were simulated in the context of this study so, ethical approval is not required.

Authors Contribution: Sebahat Gören-Investigation, methodology, software, resources, formal analysis, and writing-original draft. Hakan Kara-Investigation, methodology, visualization, resources, writing. Başak Erdem Kara-Methodology, software, visualization, formal analysis and writing. Hülya Kelecioğlu-Investigation, methodology, supervision, validation.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-127. https://doi.org/10.1177/014662168901300201

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length cats provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*(1), 1-18. https://doi.org/10.7333/1212-0101001

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item response model*. (RR 81-20). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principals and applications*. Kluwer Academic Publishers.

Jia, B., Zhang, X., & Zhu, Z. (2019). A short note on aberrant responses bias in item response theory. *Frontiers in Psychology*, *10*, 43. https://doi.org/10.3389/fpsyg.2019.00043

Liao, W., Ho, R., Yen, Y., & Cheng, H. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, *40*(10), 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–25. https://doi.org/10.1348/000711009X474502

Magis, D. (2014). On the asymptotic standard error of a class of robust estimators of ability in dichotomous item response models. *The British Journal of Mathematical and Statistical Psychology*, *67*(3), 430–450. https://doi.org/10.1111/bmsp.12027

Miller, I. & Miller, M. (2004). *John E. Freund's mathematical statistics with applications* (7th ed.). Prentice Hall.

Reckase, M., D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. Springer.

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83–101. https://doi.org/10.1177/0146621608324023

Segall, D. O. (2004). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429-438). Academic.

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. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9. https://doi.org/10.7275/wqzt-9427

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. Embretson (Ed), *New directions in psychological measurement with model-based approaches* (pp. 147-173). American Psychological Association.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 71-84. https://doi.org/10.1080/07481756.2004.11909751

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 269-281

Factors Affecting Household Expenditures on Education: A Heckman Sample Selection Application for Turkey

Abdulkerim KARAASLAN* Hasan Hüseyin TEKMANLI**

Abstract

Education is undoubtedly one of the most important elements for the development levels of countries and societies. It is also one of the essential requirements in today's world. Education is a key element of an individual's initial and later social development, therefore development across countries. Today, in most developed countries, the state spends a large amount of fund for education. Households, as well as governments, spend a lot on education. In this study, the expenses for the education of households in Turkey were discussed, and the effect of socio-demographic and economic factors on these expenditures was examined. For this purpose, the data set obtained from the 2019 Household Budget Survey published by the Turkish Statistical Institute (TURKSTAT) was analyzed with the Heckman sample selection model. According to the analysis results, the fact that the head of the household is male in a family, his age, and being married had a negative effect on education expenditures. Education expenses were positively affected by household income, population, the number of technological devices, and the employment of the head of the family. In addition, the structure of the household, the presence of individuals who smoke, go to the cinema, and do sports also affected household education expenditures.

Keywords: Household budget survey, education expenditure, Heckman sample selection, Turkey

Introduction

Education is one of the essential factors of development, and the quality of education is the key determinant of the speed at which a region achieves economic growth. Education can promote sustainable economic development through various channels, such as labor productivity increase, encouragement of good governance, reduction in income inequality, and assistance to the public health sector, and an investment made in education results in numerous economic and social benefits (Sun et al., 2019). Education both has an important effect on income and professional status and is considered a universal cure for economic problems such as unemployment and poverty (Andreou, 2012). Well-educated people bring along high levels of labor productivity. This also refers to more qualified workers and more ability to bring advanced technology from developed countries. The level and distribution of educational attainment also affect social outcomes such as child mortality, fertility, children's education, and income distribution (Barro & Lee, 2013).

Education can increase the value of labor force of individuals by developing their economic abilities, which can increase their wages and prevent poverty. For this reason, education is critical in reducing poverty (Song, 2012). Education provides people with the ability to improve themselves. Additionally, it increases the possibility of acquiring a profession and making progress in it (Cavus et al., 2021). Human capital theory argues that education enables individuals to acquire the knowledge and ability they need for a decent job accordingly, increases productivity and promotes economic growth (Song & Zhou, 2019). The importance of education, especially in developed countries' comprehensive and sustainable growth and general development of a nation, is generally accepted in the international community, as a result of which significant successes were obtained in the universalization of

* Assoc. Prof., Atatürk University, Faculty of Economics and Administrative Sciences, Erzurum-Turkey, akkaraaslan@atauni.edu.tr, ORCID ID: 0000-0002-1318-5978
** Ress. Assist., Atatürk University, Faculty of Economics and Administrative Sciences, Erzurum-Turkey, hasan.tekmanli@atauni.edu.tr, ORCID ID: 0000-0003-3687-6090

To cite this article:
Karaaslan, A., & Tekmanlı, H. H. (2022). Factors affecting household expenditures on education: A Heckman sample selection application for Turkey. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 269-281. https://doi.org/10.21031/epod.1015970

Received: 28.10.2021
Accepted: 25.09.2022

elementary education across the world. Government expenditure on education and private participation in education have also consistently increased over the years all over the world (Saha, 2013).

One of the most significant costs of raising a child is the investment made in his/her education. In this regard, household and state are two significant actors making investments in human capitals of children and youths. Household and government education expenditures are both a goal and a means to achieve other development goals, such as economic growth, poverty reduction, improved health status, greater equity, and reduced fertility (Mussa, 2013). The houses in developed countries attach great importance to the development of children's human capitals. Parents make much expenditure on education, thus expect this to help their children achieve more success and then improvement in their houses' living standards in the future (Wei et al., 2021).

As a country where the share of the working-age population in the total is increasing, Turkey has strong economic development potential, and it is doubtless that high level education will have a significant effect in achieving this potential. In Turkey, educational status is not promising. Although the average education period of the adult population is seven years, this is lower than the one in developed countries. By increasing compulsory education to 8 years in 1997 and then to 12 years in 2011, it was aimed to increase the education period of the labor force. However, it is still required to increase the education level, which necessitates more investment in education (Acar et al., 2016). The resources allocated to education in Turkey increased to 161.1 billion TL in 2019; this figure corresponds to approximately 17% of budget expenditures. This amount was allocated for the sub-items as follows: 11.4 billion for scholarships and education loans, 4.2 billion for mobile teaching and food aid, 1.2 billion for housing support, 1.4 billion for textbooks, 1.7 billion for supplementary education, 3.4 billion for the education of individuals with disabilities, 47.2 billion for higher education, 500 for higher education scholarships and loans (per student), 6 billion for private school support, 552 million TL for tuition exemption (TEDMEM, 2020). Although the resources allocated for education is increasing every year in Turkey, households must reserve a certain amount of their budgets for education as the resource allocated for state is not sufficient. In addition, some households allocate a certain part of their budget to education in order to make a difference in their economic and social gains, regardless of the limited resources they have. However, some socio-economic factors affecting households do not allow each household to make the desired investment in education; thus, these investments differ according to the household and the level of education to which the investment will be made (Demiroglari & Kiren Gurler, 2020). In Turkey, the share of educational expenditure in the total expenditure of households by years is shown in Table 1.

Table 1

Percentage of Household Expenditures of Education in Turkey by Year

Survey year	2011	2012	2013	2014	2015	2016	2017	2018	2019
Household Expenditures % Educational services	2.0	2.3	2.4	2.4	2.2	2.3	2.3	2.3	2.5

Source: TURKSTAT (2020)

According to Table 1, although there is no cumulative course, the lowest expenditure percentage was determined to be in 2011, while the highest expenditure percentage was reached in 2019.

In the field of education economy, it is a common practice to use an education production function to predict the effects of education by family background, parental education, educational expenses, etc. There is comprehensive literature showing that educational expenditures have significant effects on education (Acerenza & Gandelman, 2019; Andreou, 2012; Chi & Qian, 2016; Deng & Xue, 2014; Donkoh & Amikuzuno, 2011; Himaz, 2010; Huy, 2012; Jenkins et al., 2019; Kousar et al., 2017; Majumder & Mitra, 2016; Wei et al., 2021). There are a number of studies carried out regarding the educational expenditure of households in Turkey (Acar et al., 2016; Bayar & İlhan, 2016; Demiroglari

& Kiren Gurler, 2020; Kuvat & Ayvaz Kizilgol, 2020; San & Chaloupka, 2016; Susanli, 2013). These studies generally underlined the socio-demographic and economic conditions of households.

In a study conducted according to the 2003, 2007 and 2012 Household Budget Surveys, it was reported that the income elasticity of education increased over the years. The increase in the educational level of the household head has an increasing effect on educational expenditures. On the other hand, the household population has a reducing effect on educational expenditures (Acar et al., 2016).

According to the results obtained in a study conducted based on the 2002, 2010, and 2013 Household Budget Surveys, education expenditures of high-income households are not sensitive to changes in income level. In other words, richer households do not change their attitudes towards education expenditure significantly when there is a change in their income. However, elasticity is higher for poor households, which means that when there is a change in their income levels, they are very sensitive to such a change, and their education expenditures increase. On the other hand, if there is a decrease in level of income, education expenditures reduce more than the decrease in the level of income. In addition, the educational level of the household head affected education expenditures positively in these three years (Bayar & İlhan, 2016).

In another study carried out based on the 2017 Household Budget Survey, the increase in household income increased the likelihood of education expenditure. The increasing educational level of household head increased education expenditure. Moreover, the increase in the number of individuals living in the household reduced the likelihood of education expenditure (Kuvat & Ayvaz Kizilgol, 2020).

In another study conducted based on the 2017 Household Budget Survey regarding different educational levels, the household education expenditures were found to have a positively relationship with mother's education, father's education, mother's employment status, income, and number of computers owned in 2017 (Demiroglari & Kiren Gurler, 2020).

This study examines the factors that influence the education expenditures of households in Turkey. The primary purpose of this study is to identify the socio-demographic and economic factors that influence household participation in education expenditures and expenditure levels. Using the Heckman sample selection model, the most current household expenditure data supplied by TURKSTAT was evaluated for this purpose. In addition to socio-demographic and economic aspects, which are commonly identified in the literature regarding education expenditures, variables reflecting the cultural characteristics of the households are also included. In this regard, this study identifies cultural characteristics not previously discussed in the literature, in addition to socio-demographic and economic factors. In addition to existing education policies, these microeconometric research are essential for developing policies suited to the socioeconomic and demographic characteristics of families. The main hypothesis of this study is whether household factors are effective in education expenditure decisions and expenditure levels.

Method

Sample and Dataset

In this study, the data obtained by the 2019 Household Budget Survey was used, which was administered to 11521 houses by TURKSTAT between January 1, 2018 - December 31, 2018. This is the latest survey carried out regarding household expenditures in Turkey. TURKSTAT creates a data set covering one year by compiling the data of this survey, which is administered to an average of a thousand households every month. In addition to expenditure data of households, some socio-demographic and economic characteristics are also collected. In this data set, education expenditure data were not provided alone, but a single "education expenditure" data were obtained by collecting expenditure data on 13 different sub-items related to education by the researchers. In response to this dependent variable (EducExp), various characteristics of the family's head and the household itself were utilized as independent variables. Age (Age), gender (Male), education (Education), marital status (Married), and employment

(Employed) data of the head of the household are included in the study as independent variables; household income (Income/1000), household population (Household Population), household type (Alone, Couple, Couple with children), number of technological devices owned (Technology), car ownership (Car), smoking in the household (Cigarette), going to the movies (Cinema), doing sports (Sports) and paid TV subscription (Paid TV). The data set was reduced to 11301 households by removing the observations with missing and outlier values.

Analysis

In econometric research, the dependent variable may sometimes be continuous but limited. This mostly refers that while a dependent variable is observed to be positive for some of the population, it is zero for some of the population (Verbeek, 2008). The least-square method of a model containing such a dependent variable leads to biased empirical estimates (Long, 1997). Tobit model, which was brought to the literature by Tobin (1958), is frequently used for cases where the dependent variable takes a zero value. As dependent variables are operated through censorship in this model, statistical inefficiency is avoided. However, important limitations of the Tobit model include the fact that the assumptions of normality and homoscedasticity cannot be satisfied, which can lead to biased and inconsistent estimations (Greene, 2003). Moreover, the Tobit model mandates that the effects of independent variables on the dependent variable are the same for probabilities and levels. Due to these factors, the standard Tobit model has been utilized less frequently recently. In such a scenario, the sample selection model (Heckman, 1979) can be utilized to provide less restrictive, more consistent, and asymptotically efficient estimates for independently analyzing the effects of expenditure probability and levels. This model is also referred to as Type-II Tobit (Amemiya, 1985).

The Heckman sample selection model can be expressed using a notation similar to Yen and Rosinski (2008) as follows;

$$\begin{aligned} \log y &= x'\beta + v \quad \text{if} \quad z'\alpha + u > 0, \\ y &= 0 \qquad\quad \text{if} \quad z'\alpha + u \leq 0 \end{aligned} \tag{1}$$

where *y* denotes the dependent variable of the model, *x* and *z* are the independent variable sets, and β and α are the corresponding parameter vectors. Additionally, *u* and *v* represent bivariate normally distributed error terms with zero mean and a finite covariance matrix;

$$\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}\right) \tag{2}$$

where σ denotes the standard deviation of *v*; ρ represents the correlation between *u* and *v*. The standard deviation of *u* is unknown, thus it is set at unity, given that the selection outcomes are observed as binary, which means that the value is either 1 or 0. The sample likelihood function is;

$$L = \prod_{y=0}[1 - \Phi(z'\alpha)] \prod_{y>0} \Phi\left[\frac{z'\alpha + \rho(\log y - x'\beta)/\sigma}{(1-\rho^2)^{1/2}}\right] y^{-1} \frac{1}{\sigma}\phi\left(\frac{\log y - x'\beta}{\sigma}\right) \tag{3}$$

where y^{-1} is the Jacobian transformation from log *y* to *y*. Additionally, $\phi(.)$ and $\Phi(.)$ denote the standard normal probability density function (pdf) and cumulative distribution function (cdf), respectively. When errors are independent (ρ=0), (3) reduces to that of the two-part model, in which case the log-likelihood

function can be split into the parameters α and $[\beta,\sigma]'$. Therefore, estimation can be broken down to a probit model (to estimate α) using the whole sample and a linear regression of log y on x (to estimate β and σ) using only the non-limited observations (Cheah & Tan, 2014).

In addition, the dependent variable y is log-transformed to ameliorate potential issues with non-normality and heteroscedasticity of error terms (Aksoy et al., 2019). Calculating marginal effects by obtaining the conditional means of a log-transformed dependent variable may, however, produce misleading findings. This is demonstrated by Yen and Rosinski (2008), and they provide alternative formulas for the conditional mean of the dependent variable and marginal effects for a log-transformed sample selection model. According to Yen and Rosinski (2008), the conditional mean of the dependent variable y is;

$$E\left(y \mid y > 0\right) = \exp\left(x'\beta + \sigma^2/2\right)\Phi\left(z'\alpha + \rho\sigma\right)/\Phi\left(z'\alpha\right). \quad (4)$$

Hence, the marginal probability of a positive observation is;

$$Pr(y > 0) = \Phi\left(z'\alpha\right), \quad (5)$$

and the unconditional mean of y is;

$$E(y) = \exp\left(x'\beta + \sigma^2/2\right)\Phi\left(z'\alpha + \rho\sigma\right). \quad (6)$$

The marginal effects on probability, conditional mean, and unconditional mean are given by the derivatives of (5), (4), and (6), respectively. The marginal effects of probability determine the probability of participation in expenditures. For the conditional mean, the marginal effects determine the level of spending for those who participate in the expenditures. For the unconditional mean, the marginal effects indicate the expenditure level for the entire population, including both participants and nonparticipants. For statistical inference, standard errors were calculated by the delta method. Stata version 14.1 was used to estimate the log-likelihood function of the Heckman sample selection model.

When estimating a multiple regression model, care should be taken against the multicollinearity problem. Multicollinearity represents a highly linear intercorrelation between explanatory variables in a multiple regression model and leads to incorrect results of regression analyzes (Kim, 2019). To control the effect of multicollinearity, the variance inflation factor (VIF) should be calculated for each variable. A rule of thumb is that if the VIF for an independent variable is greater than 5 or 10, the multicollinearity of that variable is suspiciously high. In this case, it can either inflate or deflate the standard errors of the coefficients. As a result, the coefficients can, falsely, become significant (or insignificant). Another effect of multicollinearity is that a sign change of the coefficient in which a negative effect can become positive and vice versa (Tsagris & Pandis, 2021).

Results

The definitions, means, and standard deviations of dependent and independent variables are given in Table 2. Additionally, the VIF values of the independent variables are presented. The sample covers the data of 11301 houses, as seen in Table 2. 38.9% of these households made expenditures on education. While the mean monthly education expenditure for the whole sample was TL 131.831, the mean

monthly expenditure of the households making education expenditure was found to be TL 338.673. In addition, as all VIF values provided for the independent variables were less than 5, this shows that there was no multicollinearity problem between the independent variables (Alkan & Tekmanlı, 2021; Çebi Karaaslan, 2021).

Table 2

Variable Definitions and Sample Means

Variable	Definition	Mean	Std. Dev.	VIF
Dependent variable				
EducExp	Monthly education expenditure (TL)	131.831	630.406	
	Among spending (TL)	338.673	975.204	
	Expenditure rate (%)	0.389	0.488	
Continuous explanatory variables				
Education	Education of household's head	7.845	5.036	1.820
Age	Age of household's head	51.609	15.024	1.920
Technology	Number of technological devices in household	10.052	3.669	1.780
Income/1000	Monthly household income (1000 TL)	5.386	4.293	1.530
Household population	Number of household members	3.385	1.755	2.190
Binary explanatory variables (Yes = 1; No = 0)				
Male	Household's head is male	0.773	0.419	1.680
Married	Household's head is married	0.795	0.404	3.300
Employed	Household's head is employed	0.606	0.489	1.570
Alone	Household consists of one person	0.098	0.298	2.030
Couple	Household consists of couples	0.198	0.399	3.490
Couple and children	Household consists of couples and children	0.480	0.500	2.960
Car	Household has a car / cars	0.450	0.498	1.220
Cigarette	Members of household smoke	0.520	0.500	1.100
Cinema	Members of household go to cinema	0.091	0.288	1.190
Paid TV	Household has a paid TV	0.146	0.353	1.270
Sports	Members of household do sports	0.074	0.262	1.120
Sample size			*11301*	

At least one independent variable must be in the selection equation and not the level equation in the Heckman sample selection model. Ignoring such exclusion may result in a collinearity problem (Puhani, 2000). In this instance, education and household structures were omitted from the level equation and incorporated into the selection equation. In Table 3, the maximum likelihood estimation results of the Heckman sample selection model are provided. It is seen that the parameters are mostly statistically significant at which level of alpha ($p<.05$). The regression coefficients also produced economically logical results in terms of their direction. In addition, the estimated correlation coefficient (ρ) and the corresponding covariance term (λ) between the selection and level equations are statistically significant. This indicates the importance of selectivity correction. Statistical insignificance of the error correlation would have suggested a lack of endogenous sample selectivity and use of the two-part model instead.

Table 3

Maximum Likelihood Estimation of the Heckman Sample Selection Model

	Selection	Level
Variables	Estimate (Std. Error)	Estimate (Std. Error)
Constant	-1.137*** (0.098)	3.894*** (0.211)
Education	0.042*** (0.003)	
Age	-0.011*** (0.001)	0.015*** (0.003)
Technology	0.057*** (0.005)	0.065*** (0.01)
Income/1000	0.003 (0.004)	0.056*** (0.006)
Household population	0.176*** (0.01)	-0.163*** (0.021)
Male	-0.188*** (0.041)	-0.206** (0.085)
Married	-0.13** (0.056)	0.116 (0.095)
Employed	0.068** (0.034)	0.255*** (0.072)
Alone	-0.331*** (0.064)	
Couple	-0.441*** (0.059)	
Couple and children	0.232*** (0.04)	
Car	0.035 (0.029)	0.083 (0.059)
Cigarette	-0.108*** (0.027)	-0.031 (0.056)
Cinema	0.208*** (0.047)	0.154* (0.087)
Paid TV	0.167*** (0.04)	0.028 (0.076)
Sports	0.136*** (0.05)	0.025 (0.091)
σ		1.928*** (0.045)
ρ	-0.659*** (0.038)	
λ	-1.271*** (0.101)	
log-likelihood	-14463.38	

Note: Asymptotic standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.10

As Heckman sample selection is a nonlinear method, maximum likelihood estimation is not sufficient to interpret the coefficients exactly, and marginal effects are required. The marginal effect values are shown in Table 4 in order to discuss the effect size in addition to the effect direction of the variables used in the model. Here, the probability represents the likelihood of making educational expenditures, whereas the conditional and unconditional levels represent the effects on the average expenditure level of the spending households and the entire population, respectively. Marginal effects are interpreted as the effect of each added unit for continuous variables, while for binary variables, they are interpreted as the effect of having the relevant feature. The probability series is multiplied by 100 for convenience of presentation as a percentage.

Table 4

Marginal Effects of Explanatory Variables on Education Expenditures

	Probability*100	Conditional Level	Unconditional Level
Education	1.564*** (0.12)	7.093*** (0.709)	5.045*** (0.431)
Age	-0.411*** (0.045)	0.692* (0.411)	-0.437*** (0.168)
Technology	2.116*** (0.174)	20.263*** (1.557)	10.531*** (0.745)
Income/1000	0.119 (0.15)	9.826*** (1.144)	3.611*** (0.536)
Household population	6.516*** (0.382)	2.654 (2.832)	11.67*** (1.353)
Male	-6.95*** (1.514)	-65.567*** (13.271)	-34.248*** (5.63)
Married	-4.816** (2.064)	-2.671 (15.618)	-8.872 (7.179)
Employed	2.528** (1.252)	53.604*** (11.193)	22.796*** (4.635)
Alone	-12.212*** (2.359)	-55.373*** (11.001)	-39.384*** (7.65)
Couple	-16.273*** (2.183)	-73.792*** (11.129)	-52.485*** (7.264)
Couple and children	8.555*** (1.473)	38.792*** (7.069)	27.591*** (4.876)
Car	1.295 (1.062)	19.55** (9.025)	8.93** (3.791)
Cigarette	-3.972*** (1.009)	-23.204*** (8.553)	-14.615*** (3.651)
Cinema	7.7*** (1.748)	60.368*** (13.589)	33.678*** (6.106)
Paid TV	6.168*** (1.469)	32.531*** (11.519)	21.479*** (5.101)
Sports	5.012*** (1.845)	26.891* (14.11)	17.612*** (6.276)

Note: Asymptotic standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.10

According to the marginal effects of the model given in Table 4, one-year increase in the education of household head increased the probability of making education expenditure by 1.564 percent. One-year increase in the education of household head increased the conditional expenditure level by TL 7.093 and the unconditional expenditure level by TL 5.045. One-year increase in the age of household head decreased the probability of making education expenditure by 0.411 percent. This increased the conditional expenditure level by TL 0.692 and decreased the unconditional expenditure level by TL 0.437. One unit increase in the number of technological devices owned by a household increased the probability of making education expenditure by 2.116 percent. It also increased the conditional expenditure level by TL 20.263 and the unconditional expenditure level by TL 10.531. TL 1000 increase in the monthly disposable income of a household increased the conditional expenditure level by TL 9.826 and the unconditional expenditure level by TL 3.611. One person increases in the number of a household population increased the probability of making education expenditure by 6.516 percent. It also increased the unconditional expenditure level by TL 11.670.

The households headed by a male were found 6.95 percentage points less likely to make expenditure on education than households headed by a female. In addition, while the conditional expenditure levels were TL 65.567 lower, the unconditional expenditure levels were TL 34.248 lower. Families with a married head were 4.816 percentage points less likely to spend on education than the households with an unmarried head. Families with an employed head were 2.528 percentage points more likely to spend on education than the households with an unemployed head. In addition, while the conditional expenditure levels were TL 53.604 higher, the unconditional expenditure levels were TL 22.796 higher. Households with only one person were 12.212 percentage points less likely to spend on education than other households. Moreover, while the conditional expenditure levels were TL 55.373 lower, the unconditional expenditure levels were TL 39.384 lower. Households with a couple were 16.273 percentage points less likely to spend on education than other households. In addition, while the conditional expenditure levels were TL 73.792 lower, the unconditional expenditure levels were TL 52.485 lower. The households with a couple and children were 8.555 percentage points more likely to make education expenditure than other households. Moreover, while their conditional expenditure levels were TL 38.792 higher, their unconditional expenditure levels were TL 27.591 higher.

The conditional and unconditional expenditure levels of households with a car are 19.55 TL and 8.93 TL more than those without a car, respectively. The likelihood of the households with individuals smoking to make education expenditure was 3.972 percentage points lower than other households. Moreover, while the conditional expenditure level was TL 23.204 lower, the unconditional expenditure

level was TL 14.615 lower. The likelihood of the households with individuals going to cinema to make education expenditure was 7.7 percentage points higher than other households. Moreover, while the conditional expenditure level was TL 60.368 higher, the unconditional expenditure level was TL 33.678 higher. The households using paid TV were 6.168 percentage points more likely to make education expenditure than other households. In addition, while the conditional expenditure level was TL 32.531 higher, the unconditional expenditure level was TL 21.479 higher. The likelihood of the households with individuals doing sports to make education expenditure was 5.012 percentage points higher than other households. Moreover, while the conditional expenditure level was TL 26.891 higher, the unconditional expenditure level was TL 17.612 higher.

Discussion and Conclusion

This study examines the factors that influence the education expenditures of households in Turkey using the Heckman sample selection method. According to the findings of the investigation, the education level of the household's head has a positive effect on expenditure levels. Similarly, in a study carried out in Sri Lanka, it was concluded that households with more educated heads were in a tendency to demand more education (Himaz, 2010). In another study conducted for Turkey, it was determined that the increase in parent education resulted in an increase in education expenditures (Susanli, 2013). A study carried out in India reported that an increase in the education of a household head raised the share of education in the budget significantly (Azam & Kingdon, 2013). This is an expected result because well-educated individuals are aware of the value of education, and accordingly do not hesitate to spend a large part of their income for the education of their families (Jenkins et al., 2019).

An increase in the age of the household's head has a positive effect on conditional expenditure levels but a negative effect on unconditional expenditure levels and probability. In a similar vein, in a study carried out for the West Bengal Region of India, it was found out that increasing the age of household head affected education expenditure negatively in rural and urban areas (Majumder & Mitra, 2016). Moreover, it was reported in another study conducted regarding twelve Latin America - Caribbean countries and the USA, that increasing the age of household head was identified to have a negative effect on education expenditure (Acerenza & Gandelman, 2019).

An increase in the number of technological devices used in the household positively affects the probability of expenditure and its levels. Households must be rich so that family members can receive education. According to both life cycle and permanent income hypotheses, household expenditures are a function of not only current income but also wealth. In a study conducted in Ghana supported this by reporting that the durable goods variable, including mostly technological devices, had a positive effect on education expenditures (Donkoh & Amikuzuno, 2011). A similar conclusion is seen in another study carried out in Turkey. In this respect, it was found out that an increase in the number of computers in a household had an effect, especially in pre-university and university education levels. Parents assumed that their children would be educated better by computers (Demiroglari & Kiren Gurler, 2020).

An increase in the household income positively affects expenditure levels. In a study conducted in China, it was found out that families with higher income, well-educated and professional parents made more education expenditure (Qian & Smyth, 2011). It was reported in another study executed in Pakistan that as household income increased, education expenditures raised, too (Kousar et al., 2017). In a study conducted in Turkey, it was established that as the total expenditures of low-income households increased, their education expenditures increased at a slower rate. Therefore, it was concluded that education had become a necessity in this income group's household budget and that the quality of education is less important (Acar et al., 2016).

An increase in the number of household members has a positive effect on the probability of expenditure and the level of unconditional expenditure. It was determined in a study conducted in Nigeria that household population had a positive effect on education expenditures (Ogundari & Abdulai, 2014). Another study performed in China revealed that household size had a positive effect on education expenditures (Chi & Qian, 2016). In a study carried out in Thailand, it was found that larger households

were more likely to make education expenditure (Wongmonta & Glewwe, 2017). It is expected that a larger family and productive parents attach importance to children's education and make more education expenditure (Wei et al., 2021).

Male-headed households have lower expenditure probability and levels than female-headed households. In a study carried out in Nigeria, it was specified that the likelihood of the households headed by a female in rural and urban areas to make education expenditure was higher (Ogundari & Abdulai, 2014). In a similar vein, a study conducted in Nigeria reported that the households headed by a male were in a tendency to make less education expenditure (Jenkins et al., 2019). Households with married heads are less likely to spend on education. Similarly, it was reported in a study executed in China that single parents made more expenditure on their children's education (Chi & Qian, 2016). Households with a working head are more likely to spend on education and have a greater expenditure level.

While the probability of expenditure and its levels are lower for single-person and couple households, the expenditure probability and levels are greater for households consisting of couples with children. In a similar vein, in a study conducted in Vietnam, it was found out that the households with children enrolled in primary and secondary school were more likely to make education expenditure (Huy, 2012). This shows the effect of individual number in a household population. More importantly, this refers to the effect of having children, and especially the likelihood of the households with school-age children was higher to make education expenditure (Wongmonta & Glewwe, 2017).

The presence of smoking in the household has a negative effect on the probability of education expenditure and expenditure levels. This could be attributed to an exclusion effect. Similarly, a study conducted in Turkey between 2007 and 2011 revealed that cigarette use reduced education expenditures (San & Chaloupka, 2016). In a study conducted in Bangladesh, it was determined that tobacco-using households allocate a smaller portion of their budget to education expenditures (Husain et al., 2018). Similarly, a study conducted in Chile reported that households spending on tobacco allocated a smaller part of their budget to education expenditures (Paraje & Araya, 2018). Although automobile ownership has no effect on the probability of expenditure, it has a positive effect on expenditure levels. Additionally, the variables of having a household cinema habit, having a paid TV subscription, and having a sporting habit affect the probability of education expenditure and expenditure levels positively.

In Turkey, education is of importance both for households and the state, both of which make a significant investment in education. The investment made on education results in long-term expectations of qualified individuals, qualified labor, and a quality society. The fact that households make expenditures on education regardless of state investments reveals that education is attached importance by individuals. In this study, the factors affecting the education expenditure of households in Turkey have been discussed, and micro-level outputs have been obtained. As there were households not making expenditure on education, censored observation emerged, and the Heckman sample selection model has been used for analysis.

The analysis results have shown that the variables of age, education, gender, and marital status of household's head have statistically significant relationships with education expenditures. In addition, the variables of household type (single living individual, couples, couples and children), car ownership, number of technological devices, income, household population, presence of an individual smoking, going to cinemas, doing sports and households' paid TV have also statistically significant relationships with education expenditures. The majority of the study's findings are consistent with the existing literature (Acar et al., 2016; Acerenza & Gandelman, 2019; Azam & Kingdon, 2013; Chi & Qian, 2016; Demiroglari & Kiren Gurler, 2020; Himaz, 2010; Jenkins et al., 2019; San & Chaloupka, 2016; Susanli, 2013). At this point, the negative impact of the increasing age of the household's head on education expenditures draws attention. Certain applications (such as public service advertisements and TV programs) that may capture the attention of the elderly may be beneficial. In addition, the fact that male-headed households spend less on education is indicative of an issue. At this stage, it is evident that males are less inclined to invest in education. Particularly, the presence of certain educational materials in areas where men frequently spend time, such as coffee shops and tea houses, might increase the desire for education by spreading awareness. Another significant result is that smoking households spend less on education. This may be due to the fact that money that should be spent on education and other

essential requirements is instead spent on cigarettes. This situation, referred to as the "crowding-out effect" in the literature, is a very serious issue. In this scenario, the significance of anti-tobacco policies becomes apparent. In addition to the frequently used tax instrument and fines, a stronger emphasis on health and economic effects in audiovisual media may produce positive benefits.

This study differs from previous studies by subjecting variables such as individuals going to cinema, doing sports and paid TV, which refers to the socio-cultural structure of a household. The results of the study showed that all these variables had affected education expenditures positively. For this reason, it can be concluded that families with improved socio-cultural structures allocate more resources for education and attach more importance to education. In this regard, policymakers should take steps to enable a socio-cultural structure engaging in more reading, research and that is sports-oriented. This is a step that should be taken for a society allocating more resources for education and attaching more importance to education. This may contribute to laying the foundation of a society in which education is more valued.

This study is based on 2019 data for Turkey; the findings cannot be generalized to other time periods or populations. It is anticipated that the findings will shed light on the policies to be implemented. In the future, multivariate models can be developed by categorizing the sub-items for which education expenditure data is obtained. In addition, censored regression models based on panel data can be developed for training expenditures, along with the preparation of suitable data sets that take the time dimension into account.

Declarations

Author contribution: Abdulkerim KARAASLAN-Conceptualization, investigation, methodology, data curation, supervision, writing - review & editing. Hasan Hüseyin TEKMANLI-Conceptualization, methodology, writing - original draft, formal analysis, visualization.

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

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

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

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

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

References

Acar, E. Ö., Günalp, B., & Cilasun, S. M. (2016). An empirical analysis of household education expenditures in Turkey. *International Journal of Educational Development, 51*, 23-35. https://doi.org/10.1016/j.ijedudev.2016.03.007

Acerenza, S., & Gandelman, N. (2019). Household education spending in Latin America and the Caribbean: Evidence from income and expenditure surveys. *Education Finance and Policy, 14*(1), 61-87. https://doi.org/10.1162/edfp_a_00241

Aksoy, A., Bilgic, A., Yen, S. T., & Urak, F. (2019). Determinants of household alcohol and tobacco expenditures in Turkey. *Journal of Family and Economic Issues, 40*(4), 609-622. https://doi.org/10.1007/s10834-019-09619-1

Alkan, Ö., & Tekmanlı, H. H. (2021). Determination of the factors affecting sexual violence against women in Turkey: A population-based analysis. *BMC Women's Health, 21*(1), 188. https://doi.org/10.1186/s12905-021-01333-1

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Andreou, S. N. (2012). Analysis of household expenditure on education in Cyprus. *Cyprus Economic Policy Review, 6*(2), 17-38.

Azam, M., & Kingdon, G. G. (2013). Are girls the fairer sex in India? Revisiting intra-household allocation of education expenditure. *World Development, 42*, 143-164. https://doi.org/10.1016/j.worlddev.2012.09.003

Barro, R. J., & Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics, 104*, 184-198. https://doi.org/10.1016/j.jdeveco.2012.10.001

Bayar, A. A., & İlhan, B. Y. (2016). Determinants of household education expenditures: Do poor spend less on education. *Topics in Middle Eastern and North African Economies, 18*(1), 83-111.

Cavus, M., Kilinc, B. K., Yazici, B., Tekeli, S., Gunsoy, G., Gunsoy, B., & Karaduman, C. (2021). Modeling the contribution of distance education to students' preparation for the professions. *Turkish Online Journal of Distance Education, 22*(1), 106-119. https://doi.org/10.17718/tojde.849890

Cheah, Y. K., & Tan, A. K. (2014). Determinants of leisure-time physical activity: Evidence from Malaysia. *The Singapore Economic Review, 59*(02), 1450017. https://doi:10.1142/s0217590814500179

Chi, W., & Qian, X. (2016). Human capital investment in children: An empirical study of household child education expenditure in China, 2007 and 2011. *China Economic Review, 37*, 52-65. https://doi.org/10.1016/j.chieco.2015.11.008

Çebi Karaaslan, K. (2021). Analysis of factors affecting individuals' sources of happiness with multinomial logistic model. *Journal of Measurement and Evaluation in Education and Psychology, 12*(3), 286-302. https://doi.org/10.21031/epod.925631

Demiroglari, S., & Kiren Gurler, O. (2020). Determinants of household education expenditures by education level: The case of Turkey. *International Journal of Contemporary Economics and Administrative Sciences, 10*(1), 235-258. https://doi.org/10.5281/zenodo.3940537

Deng, Q., & Xue, J. (2014). Multivariate tobit system estimation of education expenditure in urban China. *Singapore Economic Review, 59*(1), 14. https://doi.org/10.1142/s0217590814500052

Donkoh, S. A., & Amikuzuno, J. A. (2011). The determinants of household education expenditure in Ghana. *Educational Research and Review, 6*(8), 570–579. http://41.66.217.101/handle/123456789/2080

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161. https://doi.org/10.2307/1912352

Himaz, R. (2010). Intrahousehold allocation of education expenditure: The case of Sri Lanka. *Economic Development and Cultural Change, 58*(2), 231-258. https://doi.org/10.1086/648187

Husain, M. J., Datta, B. K., Virk-Baker, M. K., Parascandola, M., & Khondker, B. H. (2018). The crowding-out effect of tobacco expenditure on household spending patterns in Bangladesh. *PLOS ONE, 13*(10), e0205120. https://doi.org/10.1371/journal.pone.0205120

Huy, V. Q. (2012). Determinants of educational expenditure in Vietnam. *International Journal of Applied Economics, 9*(1), 59-72.

Jenkins, G. P., Amala Anyabolu, H., & Bahramian, P. (2019). Family decision-making for educational expenditure: New evidence from survey data for Nigeria. *Applied Economics, 51*(52), 5663-5673. https://doi.org/10.1080/00036846.2019.1616075

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology, 72*(6), 558-569. https://doi.org/10.4097/kja.19087

Kousar, R., Sadaf, T., Makhdum, M. S. A., & Ijaz, A. (2017). Determinants of household's education and nutrition spending A gender-based empirical analysis. *Humanomics, 33*(4), 470-483. https://doi.org/10.1108/h-06-2016-0050

Kuvat, O., & Ayvaz Kizilgol, O. (2020). An analysis of out of pocket education expenditures in Turkey: Logit and Tobit models. *Ege Academic Review, 20*(3), 231-244. https://doi.org/10.21121/eab.795986

Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (vol. 7). Sage.

Majumder, A., & Mitra, C. (2016). Gender bias in household education expenditure: The case of West Bengal. *Indian Growth and Development Review, 9*(2), 129-150. https://doi.org/10.1108/IGDR-04-2016-0018

Mussa, R. (2013). Rural–urban differences in parental spending on children's primary education in Malawi. *Development Southern Africa, 30*(6), 789-811. https://doi.org/10.1080/0376835X.2013.859066

Ogundari, K., & Abdulai, A. (2014). Determinants of household's education and healthcare spending in Nigeria: Evidence from survey data. *African Development Review, 26*(1), 1-14. https://doi.org/10.1111/1467-8268.12060

Paraje, G., & Araya, D. (2018). Relationship between smoking and health and education spending in Chile. *Tobacco Control, 27*(5), 560-567. https://doi.org/10.1136/tobaccocontrol-2017-053857

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys, 14*(1), 53-68. https://doi.org/10.1111/1467-6419.00104

Qian, J. X., & Smyth, R. (2011). Educational expenditure in urban China: Income effects, family characteristics and the demand for domestic and overseas education. *Applied Economics, 43*(24), 3379-3394. https://doi.org/10.1080/00036841003636292

Saha, A. (2013). An assessment of gender discrimination in household expenditure on education in India. *Oxford Development Studies, 41*(2), 220-238. https://doi.org/10.1080/13600818.2013.786694

San, S., & Chaloupka, F. J. (2016). The impact of tobacco expenditures on spending within Turkish households. *Tobacco Control, 25*(5), 558–563. https://doi.org/10.1136/tobaccocontrol-2014-052000

Song, Y. (2012). Poverty reduction in China: The contribution of popularizing primary education. *China and World Economy, 20*(1), 105-122. https://doi.org/10.1111/j.1749-124X.2012.01275.x

Song, Y., & Zhou, G. (2019). Inequality of opportunity and household education expenditures: Evidence from panel data in China. *China Economic Review, 55*, 85-98. https://doi.org/10.1016/j.chieco.2019.03.002

Sun, H.-p., Sun, W.-f., Geng, Y., Yang, X., & Edziah, B. K. (2019). How does natural resource dependence affect public education spending? *Environmental Science and Pollution Research, 26*(4), 3666-3674. https://doi.org/10.1007/s11356-018-3853-6

Susanli, Z. B. (2013). *Gender and household education expenditure in Turkey*. Eurasian Business & Economics Soc. https://hdl.handle.net/11729/601

TEDMEM. (2020). *2019 Eğitim değerlendirme raporu*. https://tedmem.org/download/2019-egitim-degerlendirme-raporu?wpdmdl=3403&refresh=6057d14ac60581616367946

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*(1), 24-36. https://doi.org/10.2307/1907382

Tsagris, M., & Pandis, N. (2021). Multicollinearity. *American Journal of Orthodontics and Dentofacial Orthopedics, 159*(5), 695-696. https://doi.org/10.1016/j.ajodo.2021.02.005

TURKSTAT (2020). *Hanehalkı tüketim harcaması, 2019*. Retrieved 22.03.2021, from https://data.tuik.gov.tr/Bulten/Index?p=Hanehalki-Tuketim-Harcamasi-2019-33593#:~:text=Hanehalk%C4%B1%20b%C3%BCt%C3%A7e%20ara%C5%9Ft%C4%B1rmas%C4%B1n%C4%B1n%202019%20y%C4%B1l%C4%B1,5%20ile%20ula%C5%9Ft%C4%B1rma%20harcamalar%C4%B1%20ald%C4%B1.

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.

Wei, H., Guo, R., Sun, H., & Wang, N. (2021). Household leverage and education expenditure: The role of household investment. *Finance Research Letters, 38*, 101837. https://doi.org/10.1016/j.frl.2020.101837

Wongmonta, S., & Glewwe, P. (2017). An analysis of gender differences in household education expenditure: The case of Thailand. *Education Economics, 25*(2), 183-204. https://doi.org/10.1080/09645292.2016.1168363

Yen, S. T., & Rosinski, J. (2008). On the marginal effects of variables in the log-transformed sample selection models. *Economics Letters, 100*(1), 4-8. https://doi.org/10.1016/j.econlet.2007.10.019