
Eđitimde ve Psikolojide Ölçme ve Deęerlendirme Dergisi

Journal of Measurement
and Evaluation in
Education and Psychology

ISSN:1309-6575

Güz 2021
Autumn 2021

Cilt: 12- Sayı: 3
Volume: 12- Issue: 3



Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Sahibi

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Derneği (EPODDER)

Owner

The Association of Measurement and Evaluation in Education and Psychology (EPODDER)

Onursal Editör

Prof. Dr. Selahattin GELBAL

Honorary Editor

Prof. Dr. Selahattin GELBAL

Baş Editör

Prof. Dr. Nuri DOĞAN

Editor-in-Chief

Prof. Dr. Nuri DOĞAN

Editörler

Doç. Dr. Murat Doğan ŞAHİN
Dr. Öğr. Üyesi Eren Halil ÖZBERK
Dr. Arş. Gör. İbrahim UYSAL

Editors

Assoc. Prof. Dr. Murat Doğan ŞAHİN
Assist. Prof. Dr. Eren Halil ÖZBERK
Res. Assist. Dr. İbrahim UYSAL

Yayın Kurulu

Prof. Dr. Cindy M. WALKER
Prof. Dr. Hakan Yavuz ATAR
Prof. Dr. Neşe GÜLER
Prof. Dr. Terry A. ACKERMAN
Doç. Dr. Celal Deha DOĞAN
Doç. Dr. Hakan KOĞAR
Doç. Dr. Hamide Deniz GÜLLEROĞLU
Doç. Dr. Kübra ATALAY KABASAKAL
Doç. Dr. Nagihan BOZTUNÇ ÖZTÜRK
Doç. Dr. N. Bilge BAŞUSTA
Doç. Dr. Okan BULUT
Dr. Öğr. Üyesi Derya ÇAKICI ESER
Dr. Öğr. Üyesi Derya ÇOBANOĞLU AKTAN
Dr. Öğr. Üyesi Mehmet KAPLAN

Editorial Board

Prof. Dr. Cindy M. WALKER
Prof. Dr. Hakan Yavuz ATAR
Prof. Dr. Neşe GÜLER
Prof. Dr. Terry A. ACKERMAN
Assoc. Prof. Dr. Celal Deha DOĞAN
Assoc. Prof. Dr. Hakan KOĞAR
Assoc. Prof. Dr. Hamide Deniz GÜLLEROĞLU
Assoc. Prof. Dr. Kübra ATALAY KABASAKAL
Assoc. Prof. Dr. Nagihan BOZTUNÇ ÖZTÜRK
Assoc. Prof. Dr. N. Bilge BAŞUSTA
Assoc. Prof. Dr. Okan BULUT
Assist. Prof. Dr. Derya ÇAKICI ESER
Assist. Prof. Dr. Derya ÇOBANOĞLU AKTAN
Assist. Prof. Dr. Mehmet KAPLAN

Dil Editörü

Doç. Dr. Sedat ŞEN
Dr. Arş. Gör. Ayşenur ERDEMİR
Arş. Gör. Ergün Cihat ÇORBACI
Arş. Gör. Oya ERDİNÇ AKAN

Language Reviewer

Assoc. Prof. Dr. Sedat ŞEN
Res. Assist. Dr. Ayşenur ERDEMİR
Res. Assist. Ergün Cihat ÇORBACI
Res. Assist. Oya ERDİNÇ AKAN

Mizanpaj Editörü

Arş. Gör. Ömer KAMIŞ
Arş. Gör. Sebahat GÖREN

Layout Editor

Res. Assist. Ömer KAMIŞ
Res. Assist. Sebahat GÖREN

Sekreteryası

Arş. Gör. Aybüke DOĞAÇ
Arş. Gör. Ayşe BİLİCİOĞLU

Secretarait

Res. Assist. Ayşe BİLİCİOĞLU
Res. Asist. Aybüke DOĞAÇ

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is a national 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/pub/epod

Contact

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

Dizinleme / Abstracting & Indexing

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

Hakem Kurulu / Referee Board

Abdullah Faruk KILIÇ (Adıyaman Üni.)
Ahmet Salih ŞİMŞEK (Kırşehir Ahi Evran Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Alperen YANDI (Bolu Abant İzzet Baysal Üni.)
Asiye ŞENGÜL AVŞAR (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Arife KART ARSLAN (Başkent Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar ŞAHİN SARKIN (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengü BÖRKAN (Boğaziçi Üni.)
Betül ALATLI (Balıkesir Üni.)
Betül TEKEREK (Kahramanmaraş Sütçü İmam Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Ege Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Celal Deha DOĞAN (Ankara Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Aksaray Üni.)
Çiğdem REYHANLIOĞLU (MEB)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Ordu Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni.)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Devrim ALICI (Mersin Üni.)
Devrim ERDEM (Niğde Ömer Halisdemir Üni.)
Didem KEPİR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (Gazi Üni.)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni.)
Duygu Gizem ERTOPRAK (Amasya Üni.)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Elif Kübra Demir (Ege Üni.)
Elif Özlem ARDIÇ (Trabzon Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)

Eren Can AYBEK (Pamukkale Üni.)
Eren Halil ÖZBERK (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaraş Sütçü İmam Üni.)
Ersoy KARABAY (Kırşehir Ahi Evran Üni.)
Esin TEZBAŞARAN (İstanbul Üni.)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Ezgi MOR DİRLİK (Kastamonu Üni.)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni.)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Fuat ELKONCA (Muş Alparslan Üni.)
Fulya BARIŞ PEKMEZCİ (Bozok Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca USTA (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Görkem CEYHAN (Muş Alparslan Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Güliden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice 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.)

Hakem Kurulu / Referee Board

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 SÜN BÜL (Mersin Üni.)
Özen YILDIRIM (Pamukkale Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGİN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)
Ragıp TERZİ (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 SÜN BÜL (Mersin Üni.)
Sait Çüm (MEB)
Sakine GÖÇER ŞAHİN (University of Wisconsin
Madison)
Sedat ŞEN (Harran Üni.)

Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
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

An Application of Multilevel Mixture Item Response Theory Model Sedat ŞEN, Türker TOKER	226
Item Wording Effects in Psychological Measures: Do Early Literacy Skills Matter? Hatice Çiđdem BULUT	239
Comparison of Testlet Effect on Parameter Estimates Using Different Item Response Theory Models Esin YILMAZ KOĖAR	254
Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods Zafer ÇEPNİ, Hülya KELECİOđLU	267
Analysis of Factors Affecting Individuals' Sources of Happiness with Multinomial Logistic Model Kübranur ÇEBİ KARAASLAN	286
Monitoring Student Achievement with Cognitive Diagnosis Model Levent YAKAR, Nuri DOđAN, Şenol DOST, Nazan SEZEN YÜKSEL	303

An Application of Multilevel Mixture Item Response Theory Model

Sedat ŞEN *

Türker TOKER **

Abstract

Although the mixture item response theory (IRT) models are useful for heterogeneous samples, they are not capable of handling a multilevel structure that is very common in education and causes dependency between hierarchies. Ignoring the hierarchical structure may yield less accurate results because of violation of the local independence assumption. This interdependency can be modeled straightforwardly in a multi-level framework. In this study, a large-scale data set, TEOG exam, was analyzed with a multilevel mixture IRT model to account for dependency and heterogeneity in the data set. Sixteen different multilevel models (different class solutions) were estimated using the eighth-grade mathematics data set. Model fit statistics for these 16 models suggested the CB1C4 model (one school-level and four student-level latent classes) was the best fit model. Based on CB1C4 model, the students were classified into four latent student groups and one latent school group. Parameter estimates obtained with maximum likelihood estimation were presented and interpreted. Several suggestions were made based on the results.

Key Words: Item response theory, mixture models, multilevel mixture item response theory, maximum likelihood estimation, TEOG exam.

INTRODUCTION

Item response theory (IRT; Lord & Novick, 1968) models have been commonly used by practitioners for several testing applications, including test development, item analyses, test scoring, and differential item functioning. In contrast to the classical test theory that makes analyses on total score, IRT provides the opportunity to perform analyses based on individual test items. Examinee responses to each item are typically analyzed with a range of IRT models, including one-parameter, two-parameter, and three-parameter logistic models. Several extensions of these models have been proposed for the different data conditions (van der Linden & Hambleton, 1997). Successful applications of IRT models depend on meeting their assumptions. According to Embretson and Reise (2000), two major assumptions are required for estimating item parameters with IRT; local independence and appropriate dimensionality. Local independence indicates that the responses to an item are unrelated to any other item when the person's location is controlled (de Ayala, 2009). Appropriate dimensionality indicates that the IRT model has the correct number of trait level estimates for examinees (Embretson & Reise, 2000). de Ayala (2009) states another assumption that is called functional form assumption. This simply represents whether the data follow the function specified by the model. Additional assumptions may be needed for different estimation techniques.

Another characteristic of IRT involves the indeterminacy property which refers to the independence of item parameter estimates from sample characteristics and independence of person estimates from item characteristics. This property claims that item parameter estimates of a test should not differ based on the varying populations. Thus, a single homogenous population was expected in the traditional IRT model estimations. However, there may be situations that examinees can come from different

* Assoc. Prof., Harran University, Faculty of Education, Şanlıurfa-Turkey, sedatsen@harran.edu.tr, ORCID ID: 0000-0001-6962-4960

** Assist. Prof., Uşak University, Faculty of Education, Uşak-Turkey, tokerturker@hotmail.com, ORCID ID: 0000-0002-3038-7096

To cite this article:

Şen, S., & Toker, T. (2021). An application of multilevel mixture item response theory model. *Journal of Measurement and Evaluation in Education and Psychology*, 12(3), 226-238. doi: 10.21031/epod.893149

Received: 8.03.2021
Accepted: 23.07.2021

subpopulations (Lubke & Muthén, 2005). Assuming a single population can be considered a limitation of IRT models. Other alternatives should be used for such cases. A relatively new approach called mixture IRT was developed to examine and account for the possible subpopulations in the data (Mislevy & Verhelst, 1990; Rost, 1990). Mixture IRT models are analytically based on mixture models (McLachlan & Peel, 2000), and this mixture is achieved by combining an IRT model with a latent class analysis model. Unlike the quantitative information provided by IRT models, one of the strengths of mixture IRT models is to provide both quantitative and qualitative information about the items and examinees. In the presence of multiple populations, the application of traditional IRT models may yield biased results. In this case, the mixture IRT model would be the most appropriate approach.

Mixture IRT models have been used to investigate several psychometric issues, such as detection of differential item functioning (DIF; Cohen & Bolt, 2005), different response strategies (Mislevy & Verhelst, 1990), effects of testing accommodations (Cohen, Gregg, & Deng, 2005), and test speededness (Bolt, Cohen, & Wollack, 2002). Although the mixture IRT models are useful for heterogeneous samples, they are not capable of handling a multilevel structure, common in educational research. Ignoring the hierarchical structure may yield less accurate results because of violation of the local independence assumption (Lee, Cho, & Sterba, 2018). Multilevel models acknowledge that the data consisted of hierarchies by allowing for residual components at each level in the hierarchy. When the structure of data is nested, multilevel modeling provides more accurate estimates and inferences. In this regard, multilevel mixture IRT models (Asparouhov & Muthén, 2008; Cho & Cohen, 2010; Vermunt, 2008) were developed to account for possible dependency, such as can arise due to cluster or multistage sampling. Multilevel mixture IRT models extend the standard mixture IRT model to allow detection of nuisance dimensionality at different levels in the data. In the model, dependency is taken into account by incorporating continuous or categorical latent variables or both at the higher level. Multilevel mixture IRT models have been used in several studies including Bacci and Gnaldi (2015); Cho and Cohen (2010); Finch and Finch (2013); Jilke, Meuleman, and van de Walle (2015); Lee et al., (2018); Sen and Cohen (2020); Sen, Cohen, and Kim, (2018); Liu, Liu, and Li (2018); Li, Liu, and Liu, (2020); Tay, Diener, Drasgow, and Vermunt (2011); Varriale and Vermunt (2012); and Vermunt (2008, 2011). Except for Cho and Cohen (2010) and Sen et al. (2018), all of these studies used maximum likelihood estimation (MLE).

Purpose of the Study

Large-scale data sets (e.g., TIMSS, PISA) are typically analyzed with IRT models. Recently, researchers have started to analyze such data sets using mixture IRT models to account for the heterogeneous structure underlying the examinee population (Choi, Alexeev, & Cohen, 2015; Sen et al., 2018). Although the use of mixture IRT models for large-scale data sets has increased recently, multilevel mixture item response models are seldom used compared to single-level mixture item response models (e.g., Liu et al., 2018). The data used in this study consist of a nested structure. Students are nested in schools, along with schools nested in districts. Research mentioned above provides useful information about estimates and inferences when data have subgroups. The purpose of this study is to illustrate the application of a multilevel mixture IRT model on a large-scale data set. In this study, we attempt to show how the multilevel mixture IRT model can be used to identify and describe characteristics of latent groups in the presence of a multilevel data structure.

METHOD

Multilevel mixture IRT modeling approach was used in this study to explain the heterogeneity behind the hierarchical data set under examination. Detailed explanations about the data set and analyses are presented below.

Participants and Data Set

37,276 eighth-grade students studying in one of the provinces of the South East region of Turkey constituted the participants of this study. The sample consists of students from 521 schools from 13 districts of that province. The number of students per school varied between 1 and 609. Thirteen schools with less than 10 students were excluded from the data set in order to prevent estimation errors for hierarchical data. Thus, the remaining 508 schools with 37,199 students were used as an effective sample size in this study. The responses of these students to the Mathematics section of TEOG (Transition from Basic Education to Secondary Education) exam in November 2016 were used. There were twenty multiple-choice questions in each of four different booklets (A, B, C, and D) in the TEOG exam. Each booklet was re-coded as 0 for incorrect and 1 for correct responses. In addition, empty answers were coded as incorrect answers. After re-coding the data set, it was prepared for multilevel analyses by creating the school IDs.

Data Analysis

The multilevel mixture IRT models were used to analyze the TEOG Mathematics data set in this study. The formula of multilevel mixture IRT model can be given as follows (Lee et al., 2018, p.4):

$$\text{logit}[P(y_{jki} = 1 | \theta_{jkg}, \theta_k, C_{jk})] = \alpha_{i.g.W} \theta_{jkg} + \alpha_{i.B} \theta_k - \beta_{ig} \tag{1}$$

where j and k ($k = 1, \dots, K$) represent respondents and clusters, respectively, C_{jk} is a categorical latent variable at the within level for a respondent j nested within a cluster k , $\alpha_{i.g.W}$ is a class-specific within-level item discrimination parameter, $\alpha_{i.B}$ is a between-level item discrimination parameter, β_{ig} is a class-specific item location parameter, θ_{jkg} is a class-specific within-level continuous latent variable and θ_k is a between-level continuous latent variable. Both of these two continuous latent variables are assumed to follow a normal distribution. A sample path diagram for two level mixture IRT model with five items is displayed in Figure 1. Interested readers are referred to Lee et al. (2018) for more details.

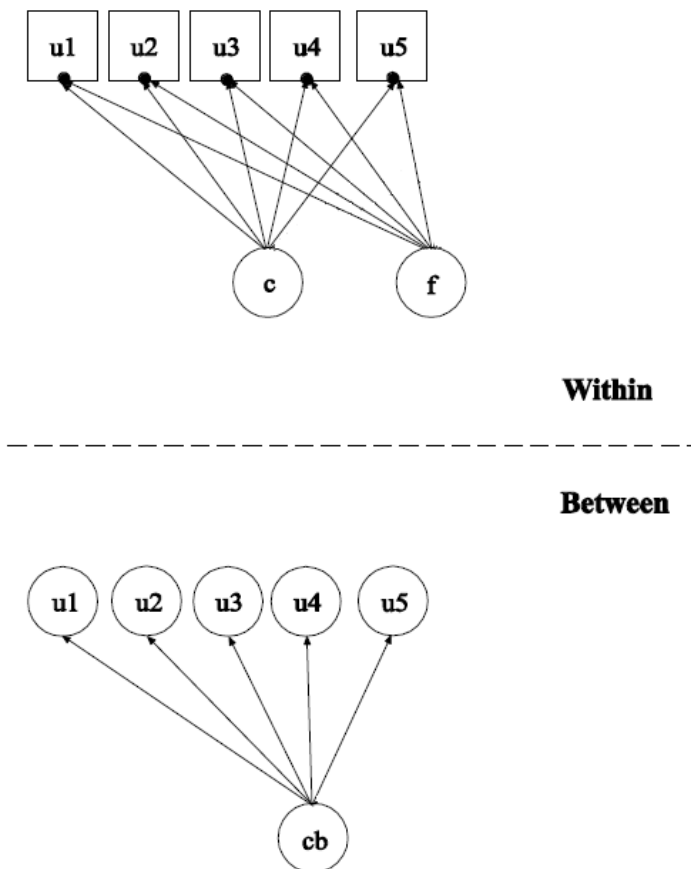


Figure 1. Diagram of the Two Level Mixture IRT Model

All analyses were conducted using Mplus 8.2 software (Muthén & Muthén, 1998-2018). Marginal maximum-likelihood estimation technique with the MLR estimator option was used for parameter estimation. For model identification, factor mean and variance were set to be 0 and 1, respectively (Muthén, 2008). The factor means in all classes were fixed to zero as the thresholds were not held equal across classes and the variances were fixed at one to set the metric of the factors. In IRT, this is usually done by fixing the factor variance to one and freeing all factor loadings. The syntax used for the final model is presented in the Appendix. TYPE = TWOLEVEL MIXTURE; ALGORITHM = INTEGRATION; options were used under ANALYSIS command in order to estimate a two level mixture IRT model. %WITHIN% and %BETWEEN% options were used to specify number of classes at each level and the relationship between items and factors under the MODEL command.

As the latent classes are unobserved and the number of classes is unknown a priori, mixture models typically follow an exploratory approach to determine the final model. Generally, it starts with a single-class model and continues by adding a class to the model until a desirable fit is obtained. Information criteria-based relative fit indices are used to determine the best-fitting model. Three information criteria indices, Akaike's Information Criterion (AIC; Akaike, 1974), Bayesian Information Criterion (BIC; Schwarz, 1978), and Sample-size Adjusted BIC (SABIC; Sclove, 1987), were used to determine the best model in terms of fit. In this study, the following formulas were used to calculate information criteria indices:

$$AIC = -2LL + 2p, \quad (2)$$

$$BIC = -2LL + \log(n) \times p, \quad (3)$$

$$SABIC = -2LL + \log\left(\frac{n+2}{24}\right) \times p. \quad (4)$$

where LL represents log-likelihood value, p denotes the number of estimated parameters and n is used for sample size. Multilevel mixture IRT models with different numbers of between and within level classes were compared in this study. The following 16 multilevel models were estimated: CB1C1 (one between level and one person level class), CB1C2, CB1C3, CB1C4, CB2C1, CB2C2, CB2C3, CB2C4, CB3C1, CB3C2, CB3C3, CB3C4, CB4C1, CB4C2, CB4C3, and CB4C4 where CB represents between-level class and C represents the within-level class. AIC, BIC, and SABIC indices were calculated for each of these models. The smallest value of each information criterion index was taken as indicating the best fitting model. Li, Cohen, Kim, and Cho (2009) and Preinerstorfer and Formann (2011) suggested that the BIC was more accurate than the AIC for model selection with single-level dichotomous mixture IRT models. In line with these studies, Sen et al. (2018) suggested that BIC was more accurate at the selection of multilevel mixture Rasch models. Therefore, BIC was used as the main index for model selection in this study.

RESULTS

As the multilevel mixture IRT model was used to analyze the data, the hierarchical structure of the data set was examined using the intra-class correlation (ICC; Raudenbush & Bryk, 2002) before conducting the analyses. A multilevel Rasch model was estimated based on the linear mixed-effects model approach using the *lmer* function (Bates & DebRoy, 2004). The ICC was .578, indicating school level can explain 57.8% of the total variance. As mentioned earlier, 16 different models were analyzed with the same data set. Model fit statistics for these 16 models are presented in Table 1. As shown in Table 1, CB1C4 (one school-level and four student-level latent classes) and CB3C4 had the smallest AIC values, CB1C4 and CB2C4 had the smallest BIC and SABIC values. Sen et al. (2018) suggested that BIC was more accurate at the selection of multilevel mixture Rasch models. Therefore, in view of these results, we conclude that the heterogeneity behind this real data can be explained by the CB1C4 model.

Table 1. Fit Statistics for Estimated Models

	LL	np	AIC	BIC	SABIC
CB1C1	-443868.808	40	887817.615	888158.577	888031.457
CB1C2	-437191.328	121	874624.657	875656.315	875271.777
CB1C3	-434580.273	182	869524.546	871076.297	870497.901
CB1C4	-433001.290	243	866488.580	868560.423	867788.169
CB2C1	-443038.394	121	886318.788	887350.447	886965.909
CB2C2	-436628.305	242	873740.609	875803.927	875034.851
CB2C3	-433632.697	363	867991.393	871086.369	869932.755
CB2C4	-432101.861	484	865171.721	869298.356	867760.204
CB3C1	-442778.648	182	885921.297	887473.048	886894.652
CB3C2	-436376.190	363	873478.379	876573.356	875419.741
CB3C3	-433708.326	544	868504.652	873142.853	871414.021
CB3C4	-431726.026	725	864902.053	871083.479	868779.429
CB4C1	-442648.253	243	885782.506	887854.349	887082.095
CB4C2	-436242.312	484	873452.624	877579.259	876041.107
CB4C3	-433412.992	725	868275.983	874457.410	872153.359
CB4C4	-432016.674	666	867365.347	871042.356	868925.808

Note. LL = Log-likelihood; np = number of parameters; AIC = Akaike's information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC; CB1C1 = one school level and one student level; CB4C4 = four school level and four student level; other model names on the first column follow the similar labeling rules.

Based on CB1C4 model, the students were classified into four latent student groups and one latent school group. Table 2 presents the final class counts and proportions for each latent class variable based on estimated posterior probabilities. Student level Class 4 is the dominant class (.499) based on the proportion of students within each latent school level class. It should be noted that the sum of the proportions reported in Table 2 equals 1.

Table 2. Final Class Counts and Proportions for Each Student Level Latent Class

Class	Count	Proportion
1	7597	.20379
2	3781	.10144
3	7302	.19589
4	18597	.49888

Item parameter estimates of the final model are presented in Table 3. Mplus output provided slope and intercept (threshold) parameters for within- and between-level separately. Thus, W and B subscripts were used to differentiate between the two levels. As shown in Table 3, slope (α) parameters were reported for each class at both levels. However, thresholds were obtained only for between level part. As explained by Şen, Cohen, and Kim (2020), IRT discrimination parameters are equal to slope parameters that are provided in Mplus output. However, item difficulty parameters can be obtained by dividing threshold values by slope values for each item. Item difficulty parameters for Class 4 appear to be positive and higher than those of other classes.

DISCUSSION and CONCLUSION

In this study, a multilevel mixture IRT model was presented and applied to a large-scale test dataset. The proposed model was a combination of an IRT model, a latent class model, and a multilevel model. Combining the advantages of these different techniques gives researchers a broad understanding of the concept. Analysis done at the individual level assumes one's standing is a product of the individual level. But individuals within a class might affect one another; thus, this makes them quantitatively comparable.

Table 3. Item Parameter Estimates of the Final Model

Item	Class 1			Class 2			Class 3			Class 4		
	$\alpha_{1,W}$	$\alpha_{1,B}$	β_1	$\alpha_{2,W}$	$\alpha_{2,B}$	β_2	$\alpha_{3,W}$	$\alpha_{3,B}$	β_3	$\alpha_{4,W}$	$\alpha_{4,B}$	β_4
1	2.406	1.185	-1.240	0.390	0.263	-0.034	2.191	0.880	-0.711	-0.851	0.025	0.723
2	2.008	1.199	-2.217	1.276	0.902	0.533	-0.166	0.456	-0.676	0.124	0.064	1.177
3	1.558	1.052	-3.589	1.655	0.684	-0.849	2.124	1.215	-2.233	0.105	0.170	0.071
4	0.508	0.819	-2.021	-0.004	0.264	0.330	0.897	0.534	0.714	0.314	0.147	1.058
5	2.345	0.987	-0.578	1.243	0.612	0.634	2.037	1.003	-0.565	-0.513	0.004	0.935
6	1.742	0.739	0.275	4.516	1.995	-1.989	0.518	0.448	0.746	0.466	0.047	0.573
7	1.132	0.566	0.102	4.954	2.244	-1.349	0.386	0.253	1.297	0.732	0.041	0.705
8	0.943	0.761	-1.687	2.031	1.035	-0.244	0.345	0.493	0.461	0.454	0.104	0.984
9	1.963	0.901	-0.985	3.017	1.301	-0.565	1.579	0.883	0.039	-0.013	0.015	1.095
10	0.922	0.729	-1.330	1.262	0.819	1.296	0.718	0.603	1.278	-0.260	-0.067	1.110
11	0.754	0.665	-0.966	0.623	0.442	1.479	1.313	0.576	0.797	0.013	0.095	1.161
12	1.223	0.752	-0.204	1.177	0.620	0.407	0.356	0.061	0.316	0.146	0.086	1.163
13	1.173	0.739	-0.910	-1.176	-0.140	0.775	0.689	0.688	1.198	0.167	0.015	1.554
14	1.237	0.740	1.211	0.176	0.061	2.073	0.454	0.227	1.681	-0.279	-0.100	1.056
15	1.259	0.500	0.435	1.473	1.012	0.656	-0.179	0.088	0.308	0.137	0.061	1.480
16	1.083	0.630	0.554	-0.061	0.162	-0.435	0.324	0.303	1.905	-0.518	-0.018	1.676
17	0.963	0.543	0.707	1.132	0.850	1.413	-0.690	-0.037	1.482	-0.164	-0.140	1.513
18	1.338	0.798	-0.351	-0.355	0.113	0.794	0.358	0.222	0.243	-0.213	0.044	1.378
19	1.266	0.688	0.910	1.851	0.766	-0.056	0.090	0.040	1.592	0.306	0.032	1.052
20	0.590	0.586	0.173	-0.011	0.161	0.878	-0.148	-0.063	1.046	-0.188	-0.006	1.469

First, an ICC value was calculated to see the ratio of the between-cluster variance to the total variance. This was done to see the proportion of the total variance in Y that is accounted for by the hierarchy. Later, different models were analyzed for model fit purposes. Using both BIC and SABIC indices one model was selected from 16 competing models.

Similar to the Vermunt (2008) study, it was found that there were differences in average latent abilities across schools. However, when a student’s ability was controlled, there were no differences in the individual item performances between schools. At this point, a detailed analysis including covariates might answer the question of why there were differences in average latent abilities across schools. Additionally, the Mplus software used in this study can estimate even more complex models; this model can be extended by adding continuous and categorical latent variables both at student and school levels while noting possible practical problems.

The proposed model can be useful for educational researchers when data are multilevel. Moreover, if there are concerns about heterogeneity in datasets when validity is the main issue for cross-cultural studies using large-scale assessment data. Also, the model can be a handful when researchers’ main interest is investigating the possible latent structures that share the same measurement model within the population. The main advantage of the proposed model is it can infer person-level measurement class along with the hierarchical class at the same time.

The multilevel mixture IRT models are becoming more popular among researchers. It is suggested that studies using the model should consider some requirements of multilevel mixture IRT models. The sample size requirement is one of the main concerns for researchers. This is mainly built on two blocks: the numbers of items and sample sizes required at each level. Simulation studies showed that $n = 5$ to 30 person-level units and 30 to 500 hierarchical-level units are required (Lukočienė, Varriale, & Vermunt, 2010).

This paper presents a general outline of multilevel mixture IRT model. The approach presented in this study has multiple theoretical and methodological advantages. Multilevel mixture IRT deals with issues of latent class models and measurement under one single model. In conclusion, the model can be used where researchers suspect latent structures within the data, when data are hierarchical, also when there is a need for cross-cultural comparisons. The results showed that these student and school-level classes are interpretable and uniquely explain how different latent ability structures spread across individuals and schools.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. doi: 10.1109/TAC.1974.1100705
- Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing.
- Bacci, S., & Gnaldi, M. (2015). A classification of university courses based on students' satisfaction: An application of a two-level mixture item response model. *Quality & Quantity*, *49*(3), 927-940. doi: 10.1007/s11135-014-0101-0
- Bates, D. M., & DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *91*(1), 1-17. doi: 10.1016/j.jmva.2004.04.013
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331-348. doi: 10.1111/j.1745-3984.2002.tb01146.x
- Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336-37. doi: 10.3102/1076998609353111
- Choi, Y. J., Alexeev, N., & Cohen, A. S. (2015). Differential item functioning analysis using a mixture 3-parameter logistic model with a covariate on the TIMSS 2007 mathematics test. *International Journal of Testing*, *15*(3), 239-253. doi: 10.1080/15305058.2015.1007241
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133-148. Retrieved from <https://www.jstor.org/stable/20461782>
- Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high-stakes mathematics test. *Learning Disabilities: Research and Practice*, *20*(4), 225-233. doi: 10.1111/j.1540-5826.2005.00138.x
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Publications.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, *73*(6), 973-993. doi: 10.1177/0013164413494776
- Jilke, S., Meuleman, B., & van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, *75*(1), 36-48. doi: 10.1111/puar.12318
- Lee, W. Y., Cho, S. J., & Sterba, S. K. (2018). Ignoring a multilevel structure in mixture item response models: Impact on parameter recovery and model selection. *Applied Psychological Measurement*, *42*(2), 136-154. doi: 10.1177/0146621617711999
- Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 499-518. doi: 10.1177/0146621608326422
- Li, M., Liu, Y., & Liu, H. (2020). Analysis of the problem-solving strategies in computer-based dynamic assessment: The extension and application of multilevel mixture IRT model. *Acta Psychologica Sinica*, *52*(4), 528-540. doi: 10.3724/SP.J.1041.2020.00528
- Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology*, *9*, Article 1372. doi: 10.3389/fpsyg.2018.01372
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison- Wesley.
- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21-39. doi: 10.1037/1082-989X.10.1.21
- Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, *40*(1), 247-283. Retrieved from <https://www.jstor.org/stable/41336886>
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215. doi: 10.1007/BF02295283
- Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In Hancock, G. R., & Samuelsen, K. M. (Eds.), *Advances in latent variable mixture models* (pp. 1-24). Charlotte, NC: Information Age Publishing, Inc.
- Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus users guide* (7th ed.). Los Angeles, CA: Author.
- Preinerstorfer, D., & Formann, A. K. (2011). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 251-262. doi: 10.1111/j.2044-8317.2011.02020.x

- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: SAGE.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271-282. doi: 10.1177/014662169001400305
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464. Retrieved from <https://www.jstor.org/stable/2958889>
- Sclove, L. S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333-343. doi: 10.1007/BF02294360
- Sen, S., & Cohen A. S. (2020). The impact of test and sample characteristics on model selection and classification accuracy in the multilevel mixture IRT model. *Frontiers in Psychology*, 11, Article 197. doi: 10.3389/fpsyg.2020.00197
- Sen, S., Cohen, A. S., & Kim, S. H. (2018). Model selection for multilevel mixture Rasch models. *Applied Psychological Measurement*, 43(4), 1-18. doi: 10.1177/0146621618779990
- Şen, S., Cohen, A., & Kim, S.-H. (2020). A short note on obtaining item parameter estimates of IRT models with Bayesian estimation in Mplus. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 11(3), 266-282. doi: 10.21031/epod.693719
- Tay, L., Diener, E., Drasgow, F., & Vermunt, J. K. (2011). Multilevel mixed-measurement IRT analysis: An explication and application to self-reported emotions across the world. *Organizational Research Methods*, 14(1), 177-207. doi: 10.1177/1094428110372674
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, 47(2), 247-275. doi: 10.1080/00273171.2012.658337
- Vermunt, J. K. (2008). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics*, 37(3&4), 285-299.
- Vermunt, J. K. (2011). Mixture models for multilevel data sets. In J. Hox & K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 59-81). New York: Routledge.

Çok Düzeyli Karma Madde Tepki Kuramı Modelinin Bir Uygulaması

Giriş

Madde tepki kuramı (MTK; Lord & Novick, 1968) modelleri, uygulayıcılar tarafından test geliştirme, madde analizi, test puanlama ve farklılaşan madde fonksiyonu dahil olmak üzere çeşitli test uygulamalarında yaygın olarak kullanılmaktadır. Toplam puan üzerinden yapılan analizlere dayanan klasik test teorisinin aksine, MTK, bireysel test maddelerine dayalı analizler yapma fırsatı sunar. Sınava girenlerin doğru-yanlış şeklinde kodlanan her bir maddeye verdiği yanıtlar tipik olarak bir parametrelili, iki parametrelili ve üç parametrelili lojistik modelleri içeren bir dizi MTK modeliyle analiz edilir. Farklı veri koşulları için bu modellerin çeşitli uzantıları önerilmiştir (van der Linden & Hambleton, 1997). MTK modellerinin uygulamalarının başarısı varsayımlarının karşılanmasına bağlıdır. Embretson ve Reise'e (2000) göre, MTK ile madde parametrelerini tahmin etmek için iki ana varsayım gereklidir; yerel bağımsızlık ve uygun boyutluluk. Yerel bağımsızlık, kişinin konumu kontrol edildiğinde bir maddeye verilen yanıtların başka herhangi bir madde ile ilgisi olmadığını gösterir (de Ayala, 2009). Uygun boyutluluk, MTK modelinin sınava giren kişiler için doğru sayıda özellik düzeyi tahminine sahip olduğunu gösterir (Embretson & Reise, 2000). de Ayala (2009), işlevsel form varsayımı olarak adlandırılan başka bir varsayım belirtir. Bu, verilerin model tarafından belirtilen işlevi takip edip etmediğini gösterir. Farklı tahmin teknikleri için ek varsayımlar gerekebilir.

MTK'nın diğer bir özelliği, madde parametre tahminlerinin örneklem özelliklerinden ve kişi tahminlerinin madde özelliklerinden bağımsızlığına atıfta bulunan değişmezliktir. Bu özellik, bir testin madde parametresi tahminlerinin değişen popülasyonlara göre farklılık göstermemesi gerektiğini iddia etmektedir. Bu nedenle, geleneksel MTK modeli tahminlerinde tek bir homojen popülasyon varsayılır. Ancak, sınava girenlerin farklı alt popülasyonlardan gelebileceği durumlar olabilir (Lubke ve Muthén,

2005). Bu durumda tek bir popülasyonun MTK modellerinin bir sınırlaması olarak kabul edilebileceği varsayılır. Bu tür durumlar için başka alternatif modeller kullanılmalıdır. Verilerdeki olası alt popülasyonları incelemek ve hesaba katmak için karma MTK adı verilen nispeten yeni bir yaklaşım geliştirilmiştir (Mislevy & Verhelst, 1990; Rost, 1990). Karma MTK modelleri analitik olarak karma modellere (McLachlan & Peel, 2000) dayalıdır ve karma MTK modeli bir MTK modeli ile bir örtük sınıf analizi modeli birleştirilerek elde edilir. MTK modelleri tarafından sağlanan nicel bilginin aksine, karma MTK modellerinin güçlü yönlerinden biri, maddeler ve sınava giren kişiler hakkında hem nicel hem de nitel bilgi sağlamasıdır. Birden fazla popülasyonun varlığında, geleneksel MTK modellerinin uygulanması yanlış sonuçlar verebilir. Bu durumda, karma MTK modeli daha uygun bir yaklaşım olacaktır.

Karma MTK modelleri, farklılaşan madde fonksiyonunun tespiti (DIF; Cohen & Bolt, 2005), farklı yanıt stratejileri (Bolt, Cohen, & Wollack, 2002; Mislevy & Verhelst, 1990) test düzenlemelerinin etkileri (Cohen, Gregg, & Deng, 2005) ve test hızının etkileri (Bolt ve diğerleri, 2002) gibi çeşitli psikometrik sorunları araştırmak için kullanılmıştır. Karma MTK modelleri heterojen örneklem için kullanışlı olsa da eğitim araştırmalarında yaygın olan çok düzeyli bir yapıyı hesaba katmamaktadır. Hiyerarşik yapıyı göz ardı etmek, düzey içi gözlemler arası bağımsızlık varsayımının ihlali nedeniyle daha yanlış sonuçlar verebilir (Lee, Cho, & Sterba, 2018). Çok düzeyli modeller, hiyerarşideki her düzeyde artık bileşenlere izin vererek verilerin hiyerarşilerden oluştuğunu kabul etmektedir. Veri yapısı iç içe olduğunda çok düzeyli modeller daha doğru tahminler ve çıkarımlar yapılmasını sağlamaktadır. Bu bağlamda, hiyerarşik veya çok düzeyli örneklemeden kaynaklanabilecek olası bağımlılığı hesaba katmak için çok düzeyli karma MTK modelleri (Asparouhov & Muthén, 2008; Cho & Cohen, 2010; Vermunt, 2008) geliştirilmiştir. Çok düzeyli karma MTK modelleri, verilerdeki farklı düzeylerde rahatsız edici boyutluluğun saptanmasına izin vermek için standart karma MTK modelini genişletir. Modelde, bağımlılık, sürekli veya kategorik örtük değişkenleri veya her ikisini üst düzeyde dahil ederek hesaba katılır. Çok düzeyli karma MTK modelleri son yıllarda birçok araştırmada kullanılmaya başlamıştır (Bacci & Gnaldi, 2015; Cho & Cohen, 2010; Finch & Finch, 2013; Jilke, Meuleman, & van de Walle, 2015; Lee ve diğerleri, 2018; Liu, Liu, & Li, 2018; Sen & Cohen 2020; Sen, Cohen, & Kim, 2018; Tay, Diener, Drasgow, & Vermunt, 2011; Varriale & Vermunt 2012; Vermunt, 2008). Cho ve Cohen (2010) ve Sen ve diğerleri (2018) dışında tüm bu çalışmalar maksimum olabilirlik tahminini (MLE) yöntemini kullanmışlardır.

Büyük ölçekli veri setleri (örneğin, TIMSS, PISA) tipik olarak MTK modelleriyle analiz edilir. Son zamanlarda araştırmacılar, incelenen popülasyonun altında yatan heterojen yapıyı hesaba katmak için bu tür veri setlerini karma MTK modelleri kullanarak analiz etmeye başlamışlardır (Choi, Alexeev, & Cohen, 2015; Sen ve diğerleri, 2018). Büyük ölçekli veri kümeleri için karma MTK modellerinin kullanımı son zamanlarda artmış olsa da çok düzeyli karma MTK modelleri, tek düzeyli karma MTK modellerine kıyasla nadiren kullanılmaktadır. Bu çalışmada kullanılan veriler iç içe bir yapıdan oluşmaktadır. Öğrenciler okullarda, okullar ise ilçeler içerisinde gruplanmaktadır. Yukarıda bahsedilen araştırmalarda, veri setleri alt gruplardan oluştuğunda çok düzeyli modellerin daha doğru tahminler ve çıkarımlar sağladığı vurgulanmaktadır. Bu çalışmanın amacı, çok düzeyli bir karma MTK modelinin hiyerarşik yapıya sahip büyük ölçekli bir veri setine uygulanmasını göstermektir. Bu çalışmada, çok düzeyli bir veri yapısının varlığında örtük sınıfların özelliklerini tanımlamak ve açıklamak için çok düzeyli karma MTK modelinin nasıl kullanılabileceğini göstermeye çalışıyoruz.

Yöntem

Bu çalışmada incelenen hiyerarşik veri setinin ardındaki heterojenliği açıklamak için çok düzeyli karma MTK modelleme yaklaşımı kullanılmıştır. Türkiye'nin Güneydoğu bölgesi illerinden birinde öğrenim gören 37,276 sekizinci sınıf öğrencisi bu çalışmanın katılımcılarını oluşturmaktadır. Örneklem, o ilin 13 ilçesinde yer alan 521 okuldaki öğrencilerden oluşmaktadır. Okul başına öğrenci sayısı 1 ile 609 arasında değişmiştir. 10'dan az öğrencisi olan 13 okul hiyerarşik veriler için tahmin hatalarını önlemek amacıyla veri setinden çıkarılmıştır. Böylece, 37,199 öğrenci ile kalan 508 okul bu çalışmada etkin örneklem büyüklüğü olarak kullanılmıştır. Örnek analizlerde bu öğrencilerin Kasım 2016'da TEOG sınavının Matematik bölümüne verdikleri yanıtlar kullanılmıştır. TEOG sınavında dört

farklı kitapçığın (A, B, C ve D) her birinde yirmi çoktan seçmeli soru vardır. Her kitapçık yanlış yanıt için 0 ve doğru yanıt için 1 olarak yeniden kodlanmıştır. Ayrıca boş cevaplar yanlış cevap olarak kodlanmıştır. Veri seti yeniden kodlandıktan sonra okul kimlikleri oluşturularak çok düzeyli analizlere hazırlanmıştır.

Bu çalışmada TEOG Matematik veri setinin analizinde çok düzeyli karma MTK modelleri kullanılmıştır. Tüm analizler Mplus 8.2 (Muthén & Muthén, 1998-2018) yazılımı kullanılarak gerçekleştirilmiştir. Parametre tahmini için marjinal maksimum olabilirlik kestirim tekniğinin sağlam versiyonu (MLR) kullanılmıştır. Model tanımlaması için faktör ortalaması ve varyansı sırasıyla 0 ve 1 olarak ayarlanmıştır (Muthén, 2008). Eşikler sınıflar arasında eşit tutulmadığından ve faktörlerin metriğini ayarlamak için varyanslar bire sabitlendiğinden, tüm sınıflardaki faktör ortalamaları sıfıra sabitlendi. MTK'da bu genellikle faktör varyansını bire sabitleyerek ve tüm faktör yüklerini serbest bırakarak yapılır.

Örtük sınıflar gözlemlenmediğinden ve sınıf sayısı önceden bilinmediğinden, karma model uygulamalarında nihai modeli belirlemek için keşfedici bir yaklaşım izlenir. Genellikle tek sınıflı bir modelle başlanır ve istenen bir uyum elde edilinceye kadar modele bir sınıf eklenerek devam edilir. En uygun modeli belirlemek için bilgi kriterlerine dayalı göreceli uyum (bilgi kriteri) indeksleri kullanılır. En iyi modeli belirlemek için Akaike'nin bilgi kriteri (AIC; Akaike, 1974), Bayesci bilgi kriteri (BIC; Schwarz, 1978) ve örneklem düzeltmeli BIC (SABIC; Sclove, 1987) olmak üzere üç bilgi kriteri indeksi kullanılmıştır.

Bu çalışmada, farklı sayıda düzey arası ve sınıf içi sınıflara sahip çok düzeyli karma MTK modelleri karşılaştırılmıştır. Hem öğrenci hem de okul düzeyindeki farklı sınıf kombinasyonlarına dayalı olarak 16 çok düzeyli model tahmin edilmiştir: CB1C1 (biri öğrenci düzeyi sınıf ve bir okul düzeyi sınıf), CB1C2, CB1C3, CB1C4, CB2C1, CB2C2, CB2C3, CB2C4, CB3C1, CB3C2, CB3C3, CB3C4, CB4C1, CB4C2, CB4C3 ve burada CB, düzeyler arası sınıfı temsil eder ve C, düzey içi sınıfı temsil eder. Bu modellerin her biri için AIC, BIC ve SABIC indeksleri hesaplanmıştır. Her bilgi kriteri indeksinin en küçük değeri, en uygun modeli gösterecek şekilde alınmıştır. Li, Cohen, Kim ve Cho (2009) ve Preinerstorfer ve Formann (2011), BIC'nin, tek seviyeli iki kategorili karma MTK modellerinin seçiminde AIC'den daha doğru olduğunu öne sürmüşlerdir. Bu çalışmalar doğrultusunda Sen ve diğerleri (2018), BIC'nin çok düzeyli karma Rasch modellerinin seçiminde diğer indekslerden daha iyi performans gösterdiğini belirtmişler. Bu nedenle, BIC bu çalışmada model seçiminde ana indeks olarak kullanılmıştır.

Sonuç ve Tartışma

Verilerin analizinde çok düzeyli karma MTK modeli kullanıldığından, analizler yapılmadan önce veri setinin hiyerarşik yapısı sınıf içi korelasyon (ICC; Raudenbush & Bryk, 2002) değeri hesaplanarak incelenmiştir. Çok düzeyli bir Rasch modeli, *lmer* fonksiyonu kullanılarak doğrusal karma etkiler modeli yaklaşımına dayalı olarak tahmin edilmiştir (Bates & Debroy, 2004). ICC değeri .578 olarak kestirilmiştir, bu da okul düzeyinin toplam varyansın %57.8'ini açıklayabileceğini gösteriyor. Daha önce de belirtildiği gibi, aynı veri seti ile 16 farklı model analiz edilmiştir. Bu 16 model için model uyum istatistikleri Tablo 1'de sunulmuştur. Tablo 1'de gösterildiği gibi, CB1C4 (1 okul düzeyinde ve 4 öğrenci düzeyinde örtük sınıf) ve CB3C4 en küçük AIC değerlerine sahipken, CB1C4 ve CB2C4 en küçük BIC ve SABIC değerlerine sahiptir. Sen ve diğerleri (2018), BIC'nin çok düzeyli karma Rasch modellerinin seçiminde daha doğru olduğunu öne sürmüştür. Bu nedenle, bu sonuçlar ışığında, bu gerçek verilerin arkasındaki heterojenliğin CB1C4 modeli ile açıklanabileceği sonucuna varılmıştır.

Bu çalışmada, büyük ölçekli bir test veri setine uygulanmış çok düzeyli bir karma MTK modeli sunulmuştur. Önerilen model, bir MTK modeli, örtük bir sınıf modeli ve çok düzeyli bir modelin bir kombinasyonudur. Bu farklı tekniklerin avantajlarını birleştirmek, araştırmacılara kavramı geniş bir şekilde anlamalarını sağlar. Bireysel düzeyde yapılan analiz, kişinin duruşunun bireysel seviyenin bir ürünü olduğunu varsayar. Ancak bir sınıftaki bireyler birbirlerini etkileyebilir, bu da onları nicel olarak karşılaştırılabilir kılar.

Bu çalışmada ilk olarak, kümeler arası varyansın toplam varyansa oranını görmek için ICC (Raudenbush & Bryk, 2002) değeri hesaplanmıştır. Bu, hiyerarşi tarafından hesaplanan toplam varyans oranını görmek için yapıldı. Daha sonra model uyumu açısından farklı modeller analiz edilmiştir. Bilgi kriteri indekslerine dayanarak, 16 alternatif model arasından en düşük uyum indeksine dayalı olan model seçilmiştir. Vermunt (2008) çalışmasına benzer şekilde, okullar arasında ortalama örtük yeteneklerde farklılıklar olduğu bulundu. Bununla birlikte, bir öğrencinin yeteneği kontrol edildiğinde, okullar arasında bireysel madde performanslarında hiçbir fark yoktu. Bu noktada, ortak değişkenleri içeren ayrıntılı bir analiz, okullar arasında ortalama örtük yeteneklerde neden farklılıklar olduğu sorusuna cevap verebilir. Ek olarak, bu çalışmada kullanılan yazılım daha karmaşık modelleri tahmin edebilir, bu model olası pratik problemlere dikkat çekerken hem öğrenci hem de okul düzeyinde sürekli ve kategorik örtük değişkenler ekleyerek genişletilebilir.

Önerilen model, veriler çok düzeyli olduğunda eğitim araştırmacıları için yararlı olabilir. Dahası, veri kümeleriyle ilgili heterojenlikle ilgili endişeler varsa, geçerlilik büyük ölçekli değerlendirme verileri kullanan kültürler arası çalışmalar için ana konu olduğunda bu modeller kullanışlı olabilir. Ayrıca, araştırmacıların asıl ilgi alanı olası örtük yapıların popülasyon içinde aynı ölçme modelini paylaştığını araştırmak olduğunda model yetersiz olabilir. Önerilen modelin temel avantajı, hiyerarşik sınıfla birlikte kişi düzeyinde ölçme sınıfını aynı anda çıkarabilmesidir.

Çok düzeyli karma MTK modelleri, araştırmacılar arasında daha popüler hale gelmektedir. Modeli kullanan çalışmaların, çok düzeyli karma MTK modellerinin bazı gereksinimlerini dikkate alması önerilir. Örneklem büyüklüğü gereksinimi, araştırmacılar için ana endişelerden biridir. Bu, temel olarak iki blok üzerine inşa edilmiştir: her seviyede gerekli olan madde sayısı ve örneklem boyutları. Simülasyon çalışmaları, $n = 5$ ila 30 kişi düzeyinde birim ve 30 ila 500 hiyerarşik düzeyde birim gerektiğini göstermiştir (Lukočienė, Varriale, & Vermunt, 2010).

Bu makale, çok düzeyli karma MTK modelinin genel bir taslağını sunar. Bu çalışmada sunulan yaklaşımın birçok teorik ve metodolojik avantajı vardır. Çok düzeyli karma MTK, örtük sınıf modelleri ve tek bir model altında ölçüm konularını ele alır. Sonuç olarak, model, araştırmacıların verilerdeki örtük yapılardan şüphelendikleri durumlarda, veriler hiyerarşik olduğunda ve kültürler arası karşılaştırmalara ihtiyaç olduğunda da kullanılabilir. Sonuçlar, bu öğrenci ve okul düzeyindeki sınıfların yorumlanabilir olduğunu ve farklı örtük yetenek yapılarının bireyler ve okullar arasında nasıl yayıldığını benzersiz bir şekilde açıkladığını göstermiştir.

Appendix. Mplus Syntax for Final Model (CB1C4)

TITLE: This is an example of a two-level mixture IRT model with one between-level class and four within-level classes

```
VARIABLE: NAMES ARE u1-u20 clus;
          USEVARIABLES = u1-u20;
          CATEGORICAL = u1-u20;
          CLASSES = cb(1) c(4);
          BETWEEN = cb;
          CLUSTER = clus;
DATA: FILE = ALLCOMBINEDMPLUS.txt;
ANALYSIS: TYPE = TWOLEVEL MIXTURE;
          ALGORITHM = INTEGRATION;
          PROCESSORS = 2;
MODEL:
    % WITHIN%
    % OVERALL%
    f BY y1-y20;

    %cb#1.c#1%
    f BY y1-y20*;
    [f@0];f@1;

    %cb#1.c#2%
    f BY y1-y20*;
    [f@0];f@1;

    %cb#1.c#3%
    f BY y1-y20*;
    [f@0];f@1;

    %cb#1.c#4%
    f BY y1-y20*;
    [f@0];f@1;

    %BETWEEN%
    % OVERALL%
    fb BY y1-y20;
    fb@1;

    %cb#1.c#1%
    fb BY y1-y20*;
    [y1$1-y20$1];
    [fb@0];

    %cb#1.c#2%
```

fb BY y1-y20*;
[y1\$1-y20\$1];
[fb@0];

%cb#1.c#3%
fb BY y1-y20*;
[y1\$1-y20\$1];
[fb@0];

%cb#1.c#4%
fb BY y1-y20*;
[y1\$1-y20\$1];
[fb@0];

SAVEDATA: file is cb1c4.txt; SAVE IS FSCORES;
OUTPUT: TECH1 TECH8;

Item Wording Effects in Psychological Measures: Do Early Literacy Skills Matter?

Hatice Çiğdem BULUT *

Abstract

While the inclusion of both positively and negatively worded items is a common practice in scales, using positively and negatively worded items together may threaten the validity of a scale. Several studies have been devoted to investigating the effects of item wording methods. The current study investigated item wording effects on the responses of 4028 Turkish fifth-grade students, who responded to the Students Confidence in Mathematics (SCM) and Students Confidence in Science (SCS) scales. The role of early literacy-related variables (i.e., early literacy activities undertaken before primary school, student performance on reading literacy tasks upon entering primary school, and duration of the children's pre-primary school attendance) on item wording effects was also examined. The investigations were conducted using confirmatory factor analysis and the correlated trait–correlated method minus one CFA- CTC(M-1) model, derived from the correlated traits-correlated methods framework. The results indicate that significant item wording effects existed in both scales. Moreover, a significant and positive effect was found in both scales relating to early literacy activities undertaken before school, but no effects were found relating to student performance on reading literacy tasks upon entering primary school or duration of the children's pre-primary school attendance. Overall, the study suggests that researchers and practitioners should consider potential effects when including both positively and negatively worded items in scales, especially scales designed for younger students.

Key Words: Item wording effects, negatively worded items, factor analytic methods, correlated traits-correlated methods, validity.

INTRODUCTION

Educational and psychological scales used in research or large-scale assessments often use a mix of positively and negatively keyed items (e.g., Kam & Meyer, 2015; Michaelides, 2019; Wang, Chen, & Jin, 2015). In the literature, including mixed-format items (i.e., negatively and positively worded items) has been common for a long time (Cronbach, 1950; Nunnally, 1978). In such scales, responses to negatively worded items are routinely recoded to align them with positively worded items so that all items follow the same direction. It is assumed that simply recoding negatively worded items will yield an equivalent opposite measure compared to positively worded items (Marsh, 1996; Nunnally, 1978). However, a considerable amount of research has revealed that negatively worded items might not function as assumed in many cases (e.g., Barnette, 2000; DiStefano & Motl, 2006; Kam & Meyer, 2015). Several studies on the phenomenon of a potential mismatch between intended and interpreted item meanings focus on “item wording effects” as the causal agents (Bolt et al., 2020; Lindwall et al., 2012; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Schmitt & Allik, 2005).

Item wording effects may be related to the respondents' age, race, reading ability, cognitive ability, and/or motivation (e.g., Michaelides, 2019; Schmitt & Allik, 2005; Weems, Onwuegbuzie, & Lustig, 2003; Yang et al., 2012). Many researchers have emphasized the importance of reading ability. In particular, negatively worded items may be more problematic when data is collected from younger respondents due to their level of language and reading skills (Peng et al., 2018). Hence, item wording effects are more likely to occur in large-scale assessments or research focusing upon younger individuals. If self-reporting scales in large-scale assessments are contaminated by variances that are attributable to negatively worded items, this is likely due to a lack of reading comprehension among

* Ph.D., Cukurova University, Faculty of Education, Adana-Turkey, hcyavuz@cu.edu.tr, ORCID ID: 0000-0003-2585-3686

To cite this article:

Bulut, H. C. (2021). Item wording effects in psychological measures: Do early literacy skills matter? *Journal of Measurement and Evaluation in Education and Psychology*, 12(3), 239-253. doi: 10.21031/epod.944067

Received: 28.05.2021

Accepted: 8.09.2021

students in the early grades. These students' interpretation of negatively worded items might lead to inaccuracy in the results, with significant implications relating to derived education policies.

Given the robust relationship between reading ability and early literacy skills, we know that students' early literacy skills contribute to their reading comprehension skills (Lonigan, Burgess, & Anthony, 2000; Storch & Whitehurst, 2002). Therefore, reading practice in early childhood should have a substantial impact on a student becoming a skilled reader (Tunmer & Hoover, 2019). Such practice also might help students to interpret negatively worded items accurately, despite their age. But the effects of early literacy skills in relation to item wording effects have not been deeply researched. We sought to address this gap by examining the relationship between item wording effects and an array of variables related to early literacy activities. We identified potential item wording effects in two different scales applied to fifth-grade students as part of an international, large-scale assessment. Then, we analyzed the relationship between early literacy skills and the discovered item wording effects. We examined whether responses to negatively worded items are different than their counterpart items and whether those responses may have differed due to the early literacy skills of the participants.

Item wording effects

When items in scales include a negative adjective, negative structure, or negative verb conjugation, these items are called "negatively worded items." Self-reporting scales often contain both positively and negatively worded items (e.g., Nunnally, 1978). The reason for this practice is to make respondents more attentive to the content of the items and to avoid response bias (i.e., response styles) in scales (e.g., Barnette, 2000). However, a considerable number of studies have repeatedly shown that including both positively and negatively worded items in a scale might distort factor structure and the inter-item correlation matrix, thereby threatening the validity and reliability of the scale (e.g., DiStefano & Motl, 2006; Kam & Meyer, 2015; Wang et al., 2015). This distortion is thought to be caused by "item wording effects," which refers to artifactual relationships and/or dimensions in a scale caused by the wording of items (Podsakoff et al., 2003).

Item wording effects occur due to the assumption that recoding negatively worded items will guarantee an equivalent opposite measure, equal to positively worded items. For example, let us assume there are two items, "I feel joyful in my school" and "I feel depressed in my school," with two response options, yes or no (this example is inspired by Spector, Van Katwyk, Brannick, and Chen's [1997] work on item direction factors). Considering the related assumption, students who respond yes to the first item should respond no to the second item. However, there might be some students who would say no to both items since those students have more neutral feelings about the school (i.e., feeling neither joyful nor depressed). Such responses could distort the contextualized factor structure of the scale. This example offers a glimpse of how item wording effects occur in scales. There are many other factors (item properties and/or respondents' characteristics) that can also cause item wording effects (Michaelides, 2019; Schmitt & Allik, 2005; Weems et al., 2003; Yang et al., 2012).

Item wording effects can also be related to language and sentence structure (e.g., word order). For example, the dimensionality of the Rosenberg Self-Esteem Scale has been examined in many language families (i.e., Indo-European and Uralic), and different results have been reported (e.g., Lindwall et al., 2012; Pullmann & Allik, 2000). While some languages follow a subject-object-verb (SOV) structure where the subject comes first, the object second, and the verb third, other languages follow a SVO structure (e.g., Turkish). Such linguistic differences play a major role in sentence comprehension (Bornkessel & Schlesewsky, 2006) and sentence processing, especially in early language development (Candan et al., 2012). Similarly, sentence negation also varies by sentence structures and language. However, researchers have not considered the relationship between differences in sentence negation and item wording effects.

Item wording effects have been found in scales of self-esteem (e.g., Tomás, Oliver, Galiana, Sancho, and Lila, 2013), anxiety (Weems et al., 2003), perceived stress (Cole, Turner & Gitche, 2019), motivation (Michaelides, 2019), personality (Kam, 2018), and social-emotional learning (Bolt, Wang, Meyer & Pier, 2020). The majority of these studies investigated the occurrence of item wording effects

in the scales using factor analytic methods. However, some of them (e.g., Bolt et al., 2020; Cole et al., 2019; Kam, 2018) utilized different methods to detect item wording effects (e.g., item response theory models or latent difference modeling). On the other hand, some studies investigated which groups of students tend to give inconsistent responses to the negatively worded items (e.g., Kam, 2018; Michaelides, 2019; Weems et al., 2003). These argue that nonalignment between positively and negatively worded items is more likely to occur with younger respondents who possess lower reading abilities or with respondents who seek higher social desirability.

Studies related to reading abilities and item wording effects have emphasized that poor reading ability leads to differential response patterns for positively and negatively worded items in scales (Gnamb & Schroeders, 2020; Weems et al., 2006). Although item wording effects can occur even in samples of graduate students or adolescent participants (Marsh, 1996; Michaelides, 2019; Weems et al., 2006), younger students' reading skills can be more problematic regarding item wording effects due to these participants' lesser development in language acquisition and reading skills (Peng et al., 2018). Michaelides (2019) indicated that the responses of linguistically less proficient respondents led to biased scores obtained from positively and negatively worded items. Given the importance of early literacy skills, as documented by the bulk of extant research (Gustafsson, Hansen, & Rosén, 2013; Melhuish, 2016; Sénéchal & LeFevre, 2002), strong early literacy skills among younger respondents might prevent problems associated with decoding and processing negatively worded items. Some studies show that early literacy skills help to improve students' reading achievement and language skills (e.g., Boyce, Innocenti, Roggman, Norman, & Ortiz, 2010; Gustafsson et al., 2013). Furthermore, these studies have reemphasized that pre-primary education and early literacy skills are very important in the long run. Consistent with this explanation, poor reading ability among younger respondents may be linked to their lesser attainment of early literacy skills. To date, the influence of younger respondents' early literacy skills has not been examined in relation to their processing of negatively worded items. This study builds on previous research that revealed the general importance of reading ability by exploring the specific importance of early literacy skills in item wording interpretation.

Purpose of the Study

This study explores the relationship between item wording effects and literacy activities by asking two research questions (RQs):

RQ 1. Do item wording effects exist in the Students Confidence in Mathematics (SCM) and Students Confidence in Science (SCS) scales?

RQ 2. Is there a relationship between item wording effects and the participants' early literacy skills?

METHOD

Sample

Data were obtained from 4028 Turkish fifth-grade students who participated in the Trends in the International Mathematics and Science Study (TIMSS) 2019 (Mullis, Martin, Foy, Kelly, & Fishbein, 2020). Of the 4028 participants, 1920 were males (47.8% of the sample). In TIMSS, a two-stage random sample design (i.e., firstly schools and then students) is used to select a representative group of students from each country (Mullis & Martin, 2017). TIMSS assesses students' learning outcomes in mathematics and science and provides trends for these subjects. TIMSS also utilizes student, teacher, parent, and school leader questionnaires to gather auxiliary information about the students' home and school contexts (Mullis et al., 2020).

Data Collection Instruments

The Students Confidence in Mathematics (SCM) and Students Confidence in Science (SCS) Scales

In the student questionnaire of TIMSS 2019, there are subject-specific self-reporting scales (i.e., Students Confidence in Mathematics and Students Confidence in Science) due to the strong relationship between the students' academic self-perception and their achievement (Mullis & Martin, 2017). In this study, the SCM and SCS were used to examine item wording effects because both scales include negatively worded items. The SCM consists of nine rating items (five are negatively worded), whereas the SCS consists of seven rating items (four are negatively worded), all measured with a four-point Likert scale (1 = agree a lot, 2 = agree, 3 = disagree, 4 = disagree a lot). Both the SCM and SCS are intended to measure a single underlying latent construct; therefore, an IRT model (i.e., the Rasch partial credit model), based on the unidimensionality assumption, was fitted to the data (Yin & Fishbein, 2020). For the Turkish fifth grade, the alpha reliability coefficients were measured at acceptable levels for the SCM and SCS, at 0.84 and 0.81, respectively (Yin & Fishbein, 2020).

Early literacy-related variables

In the home questionnaire of TIMSS 2019, parents provided information regarding their children's early literacy activities before beginning primary school, their performance on reading literacy tasks upon entering primary school, and the duration of their children's pre-primary school attendance (Mullis & Martin, 2017). In this study, we selected Early Literacy Activities Before School (ASBHELA), Early Literacy Tasks Beginning School (ASBHELT), and Student Attended Preschool (ASDHAPS) as variables. ASBHELA and ASBHELT are index scores calculated by using the Rasch partial credit model (Yin & Fishbein, 2020). The ASBHELA index is derived from items about how often parents performed a set of activities (e.g., reading books, telling stories, writing letters or words) before the child entered school; this was rated with a four-point frequency scale: often, sometimes, never, or almost never. ASBHELT is another index that is derived from items about how well the child performed a set of tasks (e.g., reading some words, reading sentences, reading a story) when the child began the first grade of primary school; this was also measured with a four-point frequency scale: very well, moderately well, not very well, not at all. Lastly, the students' preschool attendance (ASDHAPS) was derived from an item in which parents are asked if and for how long their child attended an early childhood education program; the four-point frequency scale is: 0 = "Did Not Attend" 1 = "1 Year or Less" 2 = "2 Years" 3 = "3 Years or More."

Data Analysis

For the data preparation, first, we recoded positively worded items so that higher scores on all items indicated more positive attributes. Second, the response options of ASDHAPS were combined to create a categorical variable with three levels (i.e., 0 = "Did Not Attend", 1 = "1 Year or Less", and 2 = "2 Years and More"). Then, we checked missing data and confirmed that missing values for each variable were less than 7%.

After the data preparation, the factor structures of the SCM and SCS were evaluated with confirmatory factor analysis (CFA), using Mplus 7 (Muthén & Muthén, 1998–2020). For this, we tested one-factor (Model 1), two-factor (Model 2), and bi-factor models (Model 3). Model 1 hypothesized only one latent factor (i.e., unidimensional model) for each scale as anticipated in the methodology of TIMSS 2019 for SCM and SCS. Model 2 posited two independent latent factors; while one factor was specified for positively worded items, the one factor was specified for negatively worded items. Model 3 assumed one global latent factor and two separate latent factors (i.e., one for the positively worded items and another for the negatively worded items).

To evaluate the presence of item wording effects, we used the correlated traits-correlated methods (CTCM; Marsh, 1989) framework. The CTCM framework is utilized to model multitrait-multimethod (MTMM) data (i.e., data with more than one trait and method). CTCM models enable quantifying the method effects (e.g., item wording effects) by other trait factors and variables so that researchers can

find evidence for method effects with such models (Lindwall et al., 2012). For example, we can specify two method factors (i.e., one for the positively worded items and another for the negatively worded items) in addition to trait factors (i.e., latent factor underlying the items measuring the construct of interest) to examine the validity of a scale (Yang et al., 2012). In the literature, CTCM framework has generally been used to gather convergent and discriminant validity evidence for psychological multi-dimensional constructs (i.e., traits), whose scores were obtained through the different methods. Such models consider the method and trait variance and isolate their variances so that it is possible to model traits without error and method variance (Castro-Schilo, Grimm, & Widaman, 2016).

In this framework, a method factor (i.e., for method effects/item wording effects) can be modeled with negatively worded items. As a result, the trait can be estimated free of the method effects, if there are any. Studies have used CTCM models to investigate methods effects based on negatively worded items (e.g., DiStefano & Motl, 2009; Lindwall et al., 2012; Marsh, 1996; Wu, 2008). However, such models can have convergence and admissibility problems (Fan & Lance, 2017). Therefore, we adapted a correlated trait–correlated method minus one CFA- CTC(M-1) model (Eid, 2000) (Model 4), derived from the CT-CM framework. Eid revised the CFA-CTCM model by specifying the number of method factors (M) minus 1 (e.g., only one method factor is specified either for positively or negatively worded items) to avoid identification problems. Therefore, we modeled only one method factor in this model (Model 4), associated with negatively worded items. Substantive factors (i.e., trait components) and method factors (i.e., factors associated with negatively worded items) are uncorrelated in this model. The difference between the CFA- CTC(M-1) and the CTCM models comes from including only one method but not both factors for positive and negative factors (for details, see Eid, 2000). In the last model (Model 5), we tested the method factor (i.e., item wording effects) with covariates related to early literacy skills. This model predicts the effects of these covariates on substantive factors and method factors. All models are presented in Figure 1.

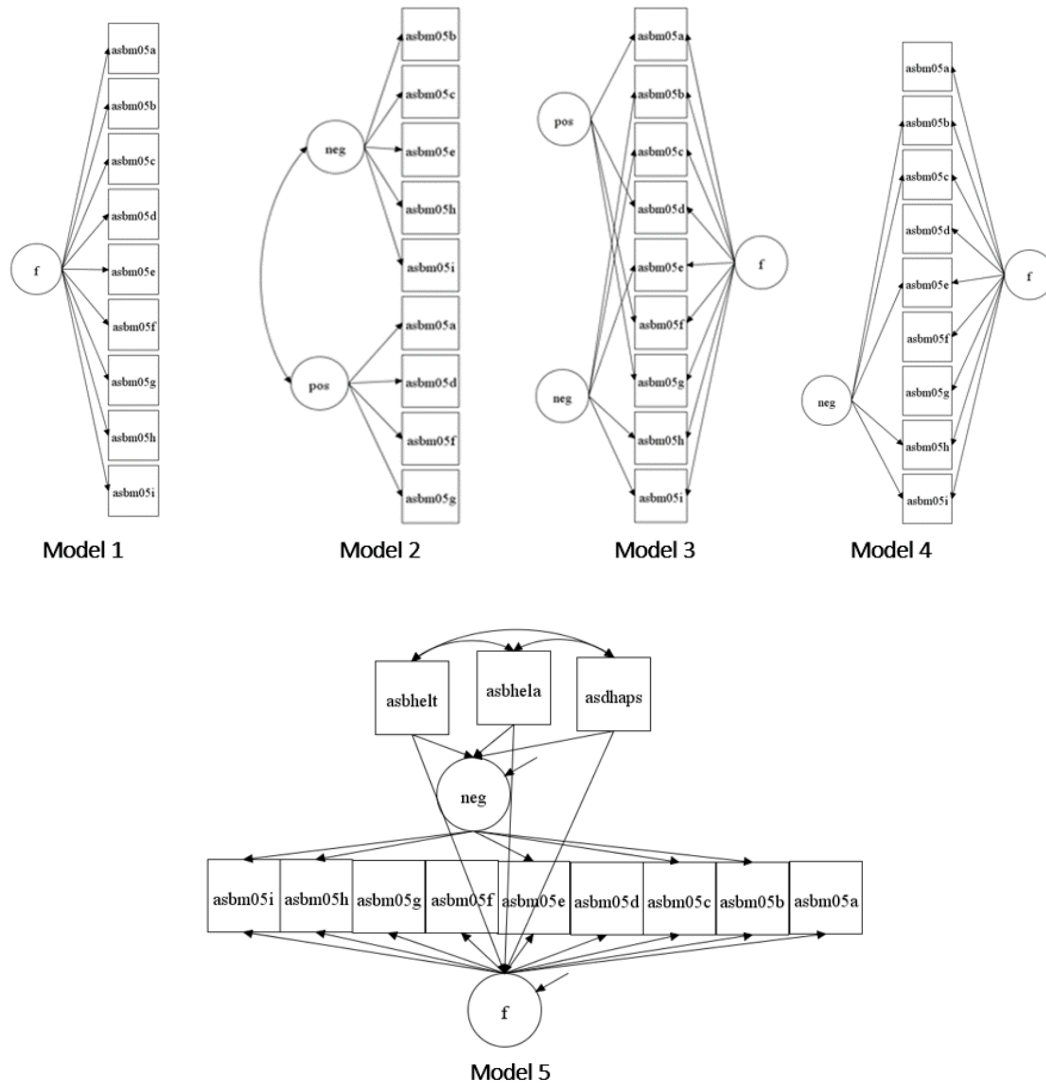


Figure 1. Path Diagrams of the Models (-for SCM)

Note: Pos = Positively worded items; Neg = Negatively worded items; f=Students' confidence in mathematics/science; Asbhelt=Early Literacy Tasks Beginning School; Asbhela= Early Literacy Activities Before School; Asdhaps= Student Attended Preschool.

We used the weighted least square mean and variance adjusted (WLSMV) to estimate the CFA models. To evaluate the models, we used several fit criteria chi-square statistics (χ^2), the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the Tucker Lewis Index (TLI). We accepted as a good fit the values of a CFI higher than .95, an RMSEA less than .05, and a TLI higher than .95, based on the recommendations of Hu and Bentler (1999).

RESULTS

Table 1 provides descriptive statistics and item-total correlations for each item. Some negatively worded items had lower mean scores than most of the positively worded items. Item-total correlations ranged from 0.36 to 0.67 ($p < .01$), indicating acceptable discrimination. However, only one item (Item 6 in the SCS = 'My teacher tells me I am good at science') fell outside the criterion (i.e., $< .40$). In addition, the standard deviations of all the negatively worded items were higher than the standard deviations of their counterpart items, indicating high variability within the negatively worded items.

Table 1. Mean, Standard Deviation, Skewness, Kurtosis, and Item-Total Correlations of the Items of the SCR and SCS

Scales	Mean	SD	Skewness	Kurtosis	Item-Total Correlations
SCM					
I usually do well in mathematics/science	3.49	0.75	-1.57	2.31	0.58
Mathematics is more difficult for me than for many of my classmates*	2.91	1.15	-0.44	-1.35	0.60
Mathematics is not one of my strengths*	3.25	1.04	-1.02	-0.41	0.67
I learn things quickly in mathematics	3.36	0.83	-1.28	1.08	0.52
Mathematics makes me nervous*	2.90	1.25	-0.51	-1.44	0.41
I am good at working out difficult mathematics problems	2.94	0.99	-0.65	-0.60	0.53
My teacher tells me I am good at mathematics	3.10	0.97	-0.86	-0.28	0.47
Mathematics is harder for me than any other subject*	2.89	1.18	-0.45	-1.37	0.66
Mathematics makes me confused*	2.79	1.17	-0.30	-1.44	0.61
SCS					
I usually do well in science	3.64	0.66	-2.10	4.64	0.48
Science is more difficult for me than for many of my classmates*	3.12	1.10	-0.77	-0.93	0.57
Science is not one of my strengths*	3.42	0.97	-1.44	0.70	0.65
I learn things quickly in science	3.49	0.79	-1.63	2.17	0.46
My teacher tells me I am good at science	3.19	0.93	-1.03	0.15	0.36
Science is harder for me than any other subject*	3.27	1.06	-1.12	-0.25	0.65
Science makes me confused*	3.16	1.09	-0.88	-0.73	0.60

* Negatively worded items. Source: Mullis, Martin, Foy, Kelly, & Fishbein (2020)

The five models presented in Figure 1 were analyzed for each scale to identify item wording effects. Table 2 presents model chi-square and fit indices for each model. Model 1 represents a one-factor model of a substantive factor (i.e., the SCM or SCS), while Model 2 represents a two-factor model, with two distinct substantive factors (i.e., the negatively worded and positively worded items of the SCM or SCS). Model 3 is a bi-factor model in which there is a general substantive factor underlying all the items and two separate two factors based on the wording of the items. On the other hand, Models 4 and 5 are CTC(M-1) models with a substantive factor (i.e., the SCM or SCS) and a method factor representing negatively worded items. Model 5 specifies the additional effect of three covariates on the method factor and substantive factors. As expected, all the models except Model 1 fit well for the data from both scales. Model 1, which did not consider item wording, provided a poor fit for the data of both scales. For both scales, Model 4 demonstrated a good fit, except for RMSEA, while Model 5 also fit the data well and was slightly better than Model 4. However, the difference between Model 4 and Model 5 is negligible. Overall, these results indicate the presence of item wording effects due to negatively worded items in the SCM and SCS scales.

Table 2. Model fit indexes for the different models for the SCM and SCS scales

SCM	x2	df	RMSEA	CFI	TLI
Model 1	2664.91	27	0.16	0.90	0.88
Model 2	539.79	26	0.07	0.98	0.97
Model 3	182.17	18	0.05	0.99	0.99
Model 4	293.94	22	0.05	0.99	0.98
Model 5	235.01	43	0.03	0.99	0.99
SCS					
Model 1	1762.56	14	0.18	0.92	0.88
Model 2	217.84	13	0.06	0.99	0.98
Model 3	42.42	7	0.03	0.99	0.99
Model 4	142.01	10	0.05	0.99	0.99
Model 5	126.49	25	0.03	0.99	0.99

Table 3 presents the results for the standardized path coefficients of the CTC(M-1) models. In Model 4 for both scales, all parameters were statistically significant, while all negatively worded items' factor loadings were higher for the method factors, except ASBM05C. As for ASBM05E, the factor loading was less than .30 for the substantive factor, while it was higher than .50 for the method factor. In Model 5, ASBHELA had a significant effect on the method factor of both scales ($p < .01$). ASBHELT and ASDHAPS did affect the method factor of the SCS scale ($p < .05$), but measures were nonsignificant for the method factor of the SCM. The size of all the significant effects may be considered low as Model 5 accounted for a low percentage of the variance in the method effects factor, with R^2 values of .03 for both scales.

Table 3. Standardized Path Coefficients for Model 4 and 5

Scales	Model 4		Model 5	
	Substantive factors	Method factors	Substantive factors	Method factors
SCM	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)	Estimate (S.E.)
ASBM05A	0.87 (0.01) **		0.87 (0.01) **	
ASBM05B	0.52 (0.02) **	0.57 (0.02) **	0.52 (0.02) **	0.56 (0.02) **
ASBM05C	0.66 (0.01) **	0.54 (0.02) **	0.66 (0.01) **	0.53 (0.02) **
ASBM05D	0.78 (0.01) **		0.77 (0.01) **	
ASBM05E	0.29 (0.02) **	0.54 (0.02) **	0.28 (0.02) **	0.53 (0.02) **
ASBM05F	0.78 (0.01) **		0.78 (0.01) **	
ASBM05G	0.71 (0.01) **		0.71 (0.01) **	
ASBM05H	0.57 (0.01) **	0.63 (0.01) **	0.57 (0.02) **	0.64 (0.01) **
ASBM05I	0.50 (0.02) **	0.63 (0.01) **	0.50 (0.02) **	0.64 (0.01) **
ASDHAPS			0.03 (0.02)	0.04 (0.02)
ASBHELA			0.03 (0.03) **	0.14 (0.03) **
ASBHELT			-0.03 (0.02)	0.02 (0.02)
SCS				
ASBS09A	0.86 (0.01) **		0.85 (0.01) **	
ASBS09B	0.44 (0.02) **	0.65 (0.02) **	0.43 (0.02) **	0.66 (0.02) **
ASBS09C	0.60 (0.02) **	0.61 (0.02) **	0.58 (0.02) **	0.62 (0.02) **
ASBS09D	0.79 (0.01) **		0.79 (0.01) **	
ASBS09E	0.65 (0.01) **		0.66 (0.01) **	
ASBS09F	0.50 (0.02) **	0.75 (0.01) **	0.50 (0.02) **	0.75 (0.01) **
ASBS09G	0.50 (0.02) **	0.62 (0.02) **	0.50 (0.02) **	0.63 (0.02) **
ASDHAPS			-0.02 (0.02)	0.06 (0.03) *
ASBHELA			0.08 (0.03) **	0.10 (0.03) **
ASBHELT			0.01 (0.03)	0.07 (0.03) *

** $p < .01$, * $p < .05$. Note: Asbhelt=Early Literacy Tasks Beginning School; Asbhela= Early Literacy Activities Before School; Asdhaps= Student Attended Preschool.

DISCUSSION and CONCLUSION

We examined the role of early literacy-related variables (i.e., early literacy activities undertaken before primary school, student performance on reading literacy tasks upon entering primary school, and duration of the children's pre-primary school attendance) on item wording effects using Turkish fifth graders' responses to the SCM and SCS scales in TIMSS 2019. Both scales were theoretically developed as a unidimensional scale and included negatively worded items. First, we applied several factor-analytic models to identify item wording effects in the scales, and then CFA- CTC(M-1) models to test them with covariates related to early literacy skills. Overall, the findings indicate that the SCM and SCS have item wording effects due to negatively worded items. However, the early literacy-related variables have insignificant or negligible effects and so cannot be used to explain the item wording effects of the SCM and SCS.

Regarding the presence of item wording effects, the results from the CFA models indicate that the inclusion of a second factor underlying the negatively worded items improved the model fit. This

suggests that anticipated factor structures for the SCM and SCS were not maintained in the Turkish sample, which indicates that negatively worded items in the SCM and SCS constituted another factor. Regardless of the subject, obtaining similar results for the confidence scales shows that students answer negatively worded items differently. The result agrees with other conclusions drawn from the literature (e.g., Michaelides, 2019; Wang et al., 2015; Yang et al., 2012). This study shows that students who participate in large-scale assessments display different tendencies when answering items based on their wording. Especially with younger age-group samples, other researchers have shown that negatively worded items might have more deleterious effects (Marsh, 1996; Michaelides, 2019; Weems et al., 2003). This might be due to the younger respondents' reading skills and different interpretations of negatively worded items (e.g., Marsh, 1996; Weems et al., 2003, 2006).

Regarding the second research question, we examined the effects of early literacy-related variables on item wording effects. We found that students' early literacy activities before school entry have significant effects on item wording effects in the SCM and SCS, but low effect sizes were found. Specifically, students engaged in early literacy activities more frequently chose higher response categories in negatively worded items than did students engaged with early literacy activities more frequently. This result indicates that students who had engaged in early literacy activities might more frequently strongly disagree in responses to negative statements compared to moderately agreeing with positively worded items. This is an interesting result and might be related to the students' personality traits (e.g., avoidance motivation, self-consciousness, and neuroticism). Quilty, Oakman, and Risko (2006) state that respondents with higher levels of avoidance motivation or neuroticism are more likely to endorse negatively worded items. Similarly, DiStefano and Molt (2005) found that other personality traits, such as reward responsiveness, fear of negative evaluation, and self-consciousness, contribute to method effects. Therefore, further investigation of the relationships between item wording effects and personality traits across younger age-group samples is recommended as a supplement to the present study. Furthermore, the seemingly counterintuitive findings may be explained by the fact that items related to the variable "students' early literacy activities before school" focus on how often instead of how deeply/successfully students engaged in these early literacy activities. In this case, it can be difficult to decide whether the frequency of doing activities or the success-rate in undertaken activities contributes more to students' early literacy skills.

Students' "performance on reading literacy tasks upon entering primary school" and "years of attending preschool" did not have significant effects on the item wording effects. This result may be due to the students' grade level, as longitudinal studies (e.g., McTigue et al., 2020; Roth, Speece, & Cooper, 2002) indicate that performance differences in early literacy may diminish or the strength of the relationship between achievement and early literacy may decline over the years, due to other sources for variation (e.g., teachers, education quality, and school). This result can additionally be supported by evidence indicating that younger respondents tend to have more problems interpreting the negative expression of a statement (Marsh, 1996; Michaelides, 2019; Weems et al., 2003, 2006). Thus, we conclude that students might interpret negatively worded items differently, regardless of their prior performance or experiences on early literacy activities. Although not a main focus in this study, Model 5 shows insignificant effects of these covariates (i.e., "students' early literacy tasks at the beginning of school" and "years of attending preschool") on the students' self-reported confidence in mathematics and science. Early literacy skills are vital to students' performance in school subjects and attitude development (Caponera, Sestito, & Russo, 2016; Petscher, 2010). However, in this study, students' early literacy skills did not lead to more confident attitudes towards mathematics and science.

Several limitations in this study must be acknowledged. First, we included a limited number of variables related to early literacy skills. Other variables (e.g., letter knowledge, vocabulary, home literacy activities, and family environment) could be included to learn more about the students' early literacy skills. Because TIMSS 2019 did not include these in their parent or student questionnaires, we could not examine the effects of such variables. Second, as data related to early literacy skills were obtained from parents, this can be problematic because self-reported data obtained from parents may be affected by the bias of social desirability. Huang (2017), for example, found that compared with teachers, parents answering the items on behalf of their children are likely to select different response categories depending on children's characteristics (e.g., gender) and parent characteristics (e.g., education level).

Therefore, in our case, parents' responses may also have been affected by these factors. Third, we did not know students' performance ratings related to their literacy skills. As a result, it is unknown whether and how variables related to early literacy skills (e.g., letter knowledge, vocabulary, home literacy activities, and family environment) affected their reading skills and the findings of this study.

Despite these limitations, this study has identified several implications for practice and future research. Firstly, we should take measures to eliminate item wording effects in the scales as much as possible in both the development and administration stages. In the development stage, researchers and practitioners should be careful when including negatively worded phrases, adjectives, and verbs within items. For instance, the item "Mathematics/Science makes me confused" was one of those which had the lowest mean scores in both scales. Therefore, "confused" can be changed to a simpler adjective that is easier for young respondents to interpret. Secondly, given the potential validity threats of item wording effects on scores obtained from scales such as the SCM and SCS, which are used in large-scale assessments, it is important to review negatively worded items in the pilot administration of the scales and to avoid administering scales that include problematic, negatively worded items – especially to relatively younger participants. Thirdly, we recommend that researchers who use data from large-scale assessments check for the presence of item wording effects. If they find evidence for this issue, then it would be beneficial for them to control these effects with a method such as CTCM or the mixed item response theory (IRT) models while estimating scale scores to avoid validity threats. Fourthly, future studies should include students' reading performance and examine how the interactions of reading performance and variables related to early literacy skills affect item wording effects in the scales. Future research also could examine the relationship between reading performance and students' interpretation of negatively worded items using larger and more representative samples and could examine whether the effects of early literacy skills on item wording effects might differ for students in earlier grades.

REFERENCES

- Barnette, J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively- worded stems. *Educational and Psychological Measurement*, 6, 361-370. doi:10.1177/00131640021970592
- Bolt, D., Wang, Y. C., Meyer, R. H., & Pier, L. (2020). An IRT mixture model for rating scale confusion associated with negatively worded items in measures of social-emotional learning. *Applied Measurement in Education*, 33(4), 331–348. doi:10.1080/08957347.2020.1789140
- Bornkessel, I., & Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychological review*, 113(4), 787. doi: 10.1037/0033-295X.113.4.787
- Boyce, L. K., Innocenti, M. S., Roggman, L. A., Norman, V. K., & Ortiz, E. (2010). Telling stories and making books: Evidence for an intervention to help parents in migrant Head Start families support their children's language and literacy. *Early Education and Development*, 21(3), 343–371. doi:10.1080/10409281003631142
- Candan, A., Küntay, A. C., Yeh, Y. C., Cheung, H., Wagner, L., & Naigles, L. R. (2012). Language and age effects in children's processing of word order. *Cognitive Development*, 27(3), 205-221. doi:10.1016/j.cogdev.2011.12.001
- Caponera, E., Sestito, P., & Russo, P. M. (2016). The influence of reading literacy on mathematics and science achievement. *The Journal of Educational Research*, 109(2), 197-204. doi: 10.1080/00220671.2014.936998
- Castro-Schilo, L., Grimm, K. J., & Widaman, K. F. (2016). Augmenting the Correlated Trait–Correlated Method Model for Multitrait–Multimethod Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(6), 798-818. doi:10.1080/10705511.2016.1214919
- Cole, K. L., Turner, R. C., & Gitche, W. D. (2019). A study of polytomous IRT methods and item wording directionality effects on perceived stress items. *Personality and Individual Differences*, 147, 63–72. doi: 10.1016/j.paid.2019.03.046
- Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 3–31. doi:10.1177/001316445001000101
- DiStefano, C., & Motl, R. W. (2006). Further investigating method effects associated with negatively worded items on self-report surveys. *Structural Equation Modeling*, 13(3), 440-464. doi: 10.1207/s15328007sem1303_6

- Dodeen, H. (2015). The effects of positively and negatively worded items on the factor structure of the UCLA loneliness scale. *Journal of Psychoeducational Assessment, 33*(3), 259-267. doi: 10.1177/0734282914548325
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241-261. doi: 10.1007/BF02294377
- Fan, Y., & Lance, C. E. (2017). A reformulated correlated trait–correlated method model for multitrait–multimethod data effectively increases convergence and admissibility rates. *Educational and psychological measurement, 77*(6), 1048-1063. doi:10.1177/0013164416677144
- Gnamb, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment, 27*(2), 404–418. doi:10.1177/1073191117746503
- Gustafsson, J.-E., Hansen, K. Y., & Rosén, M. (2013). Effects of home background on student achievement in reading, mathematics, and science at the fourth grade. In M. O. Martin and I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning*. (pp. 181- 287). Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
- Hu, L., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi: 10.1080/10705519909540118
- Huang, C. (2017). Cross-informant agreement on the child behavior checklist for youths: A meta-analysis. *Psychological reports, 120*(6), 1096-1116. doi:10.1177/0033294117717733
- Kam, C. C. S. (2018) Novel insights into item keying/valence effect using latent difference (LD) modeling analysis. *Journal of Personality Assessment, 100*(4), 389-397. doi:10.1080/00223891.2017.1369095
- Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods, 18*(3), 512–541. doi:10.1177/1094428115571894
- Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: evidence from a latent-variable longitudinal study. *Developmental psychology, 36*(5), 596. doi:10.1037/0012-1649.36.5.596
- Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*, 810–819. doi:10.1037/0022-3514.70.4.810.
- Marsh, H.W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335-361.
- McTigue, E. M., Schwippert, K., Uppstad, P. H., Lundetræ, K., & Solheim, O. J. (2020). Gender differences in early literacy: boys’ response to formal instruction. *Journal of Educational Psychology*, Advance online publication. doi:10.1037/edu0000626
- Melhuish, E. (2016). Longitudinal research and early years policy development in the UK. *International Journal of Child Care and Education Policy, 10*(1), 1-18. doi:10.1186/s40723-016-0019-1
- Michaelides, M. P. (2019). Negative keying effects in the factor structure of TIMSS 2011 motivation scales and associations with reading achievement. *Applied Measurement in Education, 32*(4), 365–378. doi: 10.1080/08957347.2019.1660349
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2019/frameworks/>
- Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2019/international-results/>
- Muthén, L. K., & Muthén, B. O. (1998-2020). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.
- Nunnally, J.C. (1978). *Psychometric Theory* (2nd edn). McGraw-Hill.
- Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., ... & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological bulletin, 144*(1), 48. doi:10.1037/bul0000124
- Petscher, Y. (2010). A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. *Journal of research in reading, 33*(4), 335-355. doi:10.1111/j.1467-9817.2009.01418.x
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. doi:10.1037/0021-9010.88.5.879
- Pullmann, H., & Allik, J. (2000). The Rosenberg Self-Esteem Scale: its dimensionality, stability and personality correlates in Estonian. *Personality and Individual Differences, 28*(4), 701-715. doi: 10.1016/S0191-8869(99)00132-4

- Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg self-esteem scale method effects. *Structural Equation Modeling*, 13(1), 99-117. doi:10.1207/s15328007sem1301_5
- R Development Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *The Journal of Educational Research*, 95(5), 259-272. doi: 10.1080/00220670209596600
- Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg self-esteem scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, 89, 623-642. doi:10.1037/0022-3514.89.4.623
- S n chal, M., & LeFevre, J. A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child development*, 73(2), 445-460. doi:10.1111/1467-8624.00417
- Spector, P. E., Van Katwyk, P. T., Brannick, M. T., & Chen, P. Y. (1997). When two factors don't reflect two constructs: How item characteristics can produce artifactual factors. *Journal of Management*, 23(5), 659-677.
- Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: evidence from a longitudinal structural model. *Developmental psychology*, 38(6), 934. doi:10.1037/0012-1649.38.6.934
- Tom s, J. M., Oliver, A., Galiana, L., Sancho, P., & Lila, M. (2013). Explaining method effects associated with negatively worded items in trait and state global and domain-specific self-esteem scales. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(2), 299-313. doi:10.1080/10705511.2013.769394
- Tunmer, W. E., & Hoover, W. A. (2019). The cognitive foundations of learning to read: A framework for preventing and remediating reading difficulties. *Australian Journal of Learning Difficulties*, 24(1), 75-93. doi:10.1080/19404158.2019.1614081
- Wang, W., Chen, H., & Jin, K. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, 75(1), 157-178. doi:10.1177/0013164414528209
- Weems, G. H., Onwuegbuzie, A. J., & Collins, K. M. (2006). The role of reading comprehension in responses to positively and negatively worded items on rating scales. *Evaluation & Research in Education*, 19(1), 3-20. doi:10.1080/09500790608668322
- Weems, G. H., Onwuegbuzie, A. J., & Lustig, D. (2003). Profiles of respondents who respond inconsistently to positively-and negatively-worded items on rating scales. *Evaluation & Research in Education*, 17(1), 45-60. doi:10.1080/14664200308668290
- Yang, Y., Chen, Y. H., Lo, W. J., & Turner, J. E. (2012). Cross-cultural evaluation of item wording effects on an attitudinal scale. *Journal of Psychoeducational Assessment*, 30(5), 509-519. doi: 10.1177/0734282911435461

Psikolojik  l eklerde Madde İfade Etkisi: Erken Okuryazarlık Becerileri Fark Yaratıyor Mu?

Giriş

Eđitim ve psikoloji alanında kullanılan  l eklerde olumlu ve olumsuz y nde ifade edilmiř maddeler birlikte bulunabilmektedir ( rn., Kam & Meyer, 2015; Michaelides, 2019; Wang ve diđ., 2015). Bunun nedeni olarak bu t r  l me ara larında olumlu y nde ifade edilmiř maddelerin yanında olumsuz y nde ifade edilmiř maddelerin yer almasının yaygın bir yaklařım olması g sterilebilir (Cronbach, 1950; Nunnally, 1978). Bu t r  l me ara larında olumsuz y nde ifade edilmiř maddeler ters kodlanarak puanlamaya katılır. Bu iřlemlerle birlikte bu maddelerin olumlu y nde ifade edilmiř maddeler gibi  alıřacağı varsayılmaktadır (Marsh, 1996; Nunnally, 1978). Fakat, alan yazındaki  alıřmalar olumsuz y nde ifade edilmiř maddelerin varsayılan řekilde iřlemediđini ortaya koymaktadır ( rn., Barnette, 2000; DiStefano & Motl, 2006; Kam & Meyer, 2015). Yapılan bazı arařtırmalar, olumsuz y nde ifade edilen maddelerin  l me aracından elde edilen puanların ge erliđini tehdit ettiđini ve g venirliđini d ř rd đ n  g stermektedir (Barnette, 2000; DiStefano & Motl, 2006; Kam & Meyer, 2015). Bunun nedeni olarak ise alan yazında madde ifade etkisi (item wording effect) olarak tanımlanan, bireylerin olumsuz y nde ifade edilmiř maddeleri farklı anlamlandırmalarından dolayı olumsuz maddelerin kendi aralarında ayrı bir fakt r oluřturması durumu g r lmektedir (DiStefano & Motl, 2006; Dodeen, 2015).

Madde ifade etkisi, olumsuz maddelerin ters kodlandıktan sonra olumlu ifade edilmiş maddeler gibi aynı yönde ve ölçme gücünde işlemediğinde ortaya çıkmaktadır. Bu duruma bir örnek verecek olursak, bir ölçekte şu iki maddenin olduğunu düşünelim: “Okulumda kendimi neşeli hissediyorum.” ve “Okulumda kendimi depresif hissediyorum.” (Örnek Spector ve diğerlerinin [1997] çalışmasından uyarlanmıştır). İlgili varsayım düşünüldüğünde ilk maddeye evet yanıtını veren bireyin ikinci maddeye hayır yanıtını vermesi beklenir. İkinci maddeye verilen yanıtlar ters kodlandığında birinci maddeye benzer şekilde yanıt örüntülerinin oluşacağı varsayılır. Fakat bazı yanıtlayıcılar okullarında kendilerini ne neşeli ne de depresif hissetmedikleri için her iki maddeye de hayır yanıtını verebilir. Bu durumda, ikinci madde ters kodlandığında veri setinde öngörülme yanıt örüntüleri ortaya çıkmaktadır. Bu tür yanıt veren bireylerin yanıtları veri setinde olduğunda ilgili ölçeğin teorik açıdan öngörülen faktör yapısı etkilenmektedir. Bu basit örnek sadece madde ifade etkisinin nasıl oluşabileceğini anlatmak için verilmiştir. Bunun yanında, madde ifade etkisinin oluşmasına yol açan birçok değişken (madde ve/veya yanıtlayıcı özellikleri) bulunmaktadır (Michaelides, 2019; Schmitt & Allik, 2005; Weems ve diğ., 2003; Yang ve diğ., 2012).

Alan yazındaki çalışmalar, ölçeklerde madde ifade etkisinin yanıtlayıcıların yaşına, kültürel özelliklerine, okuduğunu anlama becerilerine, bilişsel becerilerine ve motivasyonlarına göre ortaya çıkabileceğini göstermişlerdir (örn., Michaelides, 2019; Schmitt & Allik, 2005; Weems ve diğ., 2003; Yang ve diğ., 2012). Bu konuda yapılan çalışmalar okuduğunu anlamamanın önemini vurgulamaktadır. Özellikle, olumsuz yönde ifade edilmiş maddelerin küçük yaş gruplarına uygulanan ölçeklerde daha fazla problem yarattığı belirtilmektedir. Bunun nedeni olarak bu yaş grubundaki bireylerin dil ve okuduğunu anlama becerilerinin hala gelişim sürecinde olması gösterilmektedir (Peng ve diğ., 2018).

Okuduğunu anlama becerileri ile erken okuryazarlık becerileri arasındaki ilişki dikkate alındığında, öğrencilerin erken okuryazarlık becerilerinin okuduğunu anlama becerilerinde önemli bir rol oynadığını bilinmektedir (Lonigan ve diğ., 2000; Storch & Whitehurst, 2002). Bu nedenle, erken çocukluk döneminde okuma aktiviteleriyle ilgili daha çok tecrübe sahibi olan bireyler okuduklarını daha iyi anlamaktadırlar (Tunmer & Hoover, 2019). Buradan hareketle, bu bireylerin yaşlarına rağmen olumsuz yönde ifade edilmiş maddeleri doğru şekilde anlamlandırması beklenebilir. Bahsedilen ilişkinin önemine rağmen, alan yazında erken okuryazarlık becerilerinin olumsuz madde etkisinde bir etkisi olup olmadığı çalışılmamıştır. Bu nedenle, bu çalışmanın amacı erken okuryazarlık becerileriyle ilgili olan aktivitelerin olumsuz yönde ifade edilmiş maddeleri anlamlandırmada farklılık yaratıp yaratmadığını incelemektir. Bunun için beşinci sınıf öğrencilerine uygulanmış geniş ölçekli bir testte yer alan iki farklı ölçekte madde ifade etkisinin varlığı araştırılmıştır. Bunun yanında, bazı erken okuryazarlıkla ilgili değişkenlerin olası bu etki üzerindeki rolü incelenmiştir.

Yöntem

Bu çalışmanın örneklemini Uluslararası Matematik ve Fen Eğilimleri Araştırması (the Trends in International Mathematics and Science Study [TIMSS]) 2019’a katılmış 4028 (%47.8 erkek) beşinci sınıf Türk öğrencileri oluşturmaktadır (Mullis ve diğ., 2020). TIMSS, dört yılda bir katılımcı ülkelerin dördüncü/ beşinci ve sekizinci sınıf öğrencilerinin matematik ve fen alanlarında başarılarını belirlemeyi amaçlamaktadır. Ayrıca, TIMSS öğrencilerden, öğrencilerin öğretmenlerinden ve okul yöneticilerinden çok yönlü bilgi toplamaktadır. Bu amaçla, TIMSS başarı testleri dışında birçok ölçeği de içinde bulunduran anketleri de uygulanmaktadır.

Bu çalışmada, dörtlü Likert tipinde olan “Matematik Dersinde Kendine Güvenme” (The Students Confident in Mathematics [SCM]) ve “Fen Bilimleri Dersinde Kendine Güvenme” (The Students Confident in Science [SCS]) ölçekleri kullanılmıştır (Mullis ve diğ., 2020). SCM’de beş olumsuz ve dört olumlu yönde ifade edilmiş madde, SCS’de ise dört olumsuz ve üç olumlu yönde ifade edilmiş madde bulunmaktadır. Ölçeklerin teorik açıdan tek boyutlu olduğu ifade edilmektedir. Türk öğrencilerinin veri setlerinde, alfa güvenirlik katsayıları SCM ve SCS için sırasıyla 0.84 ve 0.81 olduğundan, ölçeklerin güvenirlik katsayıları kabul edilebilir düzeydedir (Yin & Fishbein, 2020).

TIMSS 2019’da ev anketinde ebeveynler çocuklarının erken okuryazarlıklarına ilişkin bazı soruları yanıtlamışlardır. Bu çalışmada, bu anketten “Okuldan Önce Yapılan Erken Okuryazarlık Aktiviteleri”

(ASBHELA), “Okula Başlarken Yapılan Erken Okuryazarlık Çalışmaları” (ASBHELT) ve “Öğrencilerin Okul Öncesi Eğitime Katılımı” (ASDHAPS) değişkenleri ele alınmıştır. ASBHELA ve ASBHELT Rasch kısmi puanlama modeli kullanılarak hesaplanan indeks puanlarıdır (Yin & Fishbein, 2020). ASDHAPS ise öğrencilerin okul öncesi eğitime katılıp katılmadığını, katıldıysa ne kadar katıldığını gösteren kategorik bir değişkendir.

Çalışmada verilerin analizinde öncelikle kayıp veriler incelenmiştir ve her bir değişkenin kayıp veri değerinin %7’den az olduğu görülmüştür. Sonrasında olumsuz maddeler ters kodlanmıştır. Bu amaçla öncelikle SCM ve SCS’nin faktör yapısı tek -faktör (Model 1), iki-faktör (Model 2) ve bifaktör modeli (Model 3) ile incelenmiştir. Madde ifade etkisinin varlığı ise ilişkili- özellik ilişkili yöntem (correlated traits-correlated methods-[CTCM; Marsh, 1989]) modeli kullanılarak incelenmiştir. CTCM, çoklu özellik-çoklu yöntem matrislerini modellemede kullanılmaktadır. Bu model çerçevesinde bir yöntem faktörü (yöntem etkisi/ madde ifade etkisi) modele dahil edilerek, özellikler/ gizil yapılar (traits) bu yöntemin etkisi kaldırılarak kestirilebilir. Bunun yanında, bu tür modeller yakınsama ve kabul edilebilirlik problemleri gösterebilmektedir (Fan & Lance, 2017). Bu nedenle, çalışmada ilişkili özellik- ilişkili yöntem (M-1) modeli kullanılmıştır (correlated trait-correlated method minus one CFA-CTC(M-1) model [Eid, 2000]) (Model 4). Bu modelde, olumsuz yönde ifade edilmiş maddelerin bağlandığı sadece bir yöntem faktörü tanımlanmıştır. Bu yöntem faktörü ile gizil değişkene ilişkin faktörler arasındaki korelasyon tanımlanmamıştır. Son modelde (Model 5), Model 4’te tanımlanan yöntem faktörüne ve gizil değişkene ilişkin faktöre erken okuryazarlıkla ilgili üç kovaryant değişkeni eklenmiştir. Modellerin değerlendirilmesinde ki-kare istatistiği (χ^2) ve bazı uyum indeksleri (Tucker Lewis indeksi - the Tucker Lewis Index [TLI], Karşılaştırmalı uyum indeksi- Comparative Fit Index [CFI], Ortalama hata karekök yaklaşımı- Root mean square error approximation [RMSEA]) dikkate alınmıştır. Modellerin performans kriteri olarak RMSEA’nın .05’ten düşük olması, TLI ve CFI’nın ise .95’ten büyük olması dikkate alınmıştır. Analizlerin hepsi Mplus 7 (Muthén & Muthén, 1998–2020) ve R (R Development Core Team, 2021) kullanılarak yapılmıştır.

Sonuç ve Tartışma

Bu çalışmada, teorik açıdan tek boyutlu olduğu öngörülen, beşinci sınıf Türk öğrencilere uygulanmış SCM ve SCS ölçeklerinde madde ifade etkisinin olup olmadığı incelenmiştir. Aynı zamanda, bu çalışmada erken okuryazarlık becerileriyle ilgili olan aktivitelerin, olumsuz yönde ifade edilmiş maddelerin anlamlandırmasında farklılık yaratıp yaratmadığı da araştırılmıştır. Çalışmada analiz edilen Model 1, 2, 3 ve 4’ün sonuçları ele alındığında hem SCM hem de SCR’de madde ifade etkisinin olduğu belirlenmiştir. Olumsuz yönde ifade edilmiş maddeler için ayrı tanımlanmış faktörün olduğu modeller daha iyi uyum göstermiştir. Özet olarak, öğrencilerin maddeleri ifade edilmiş yönlerine göre farklı yorumladıklarını belirtilebilir. Bu durum, alinyazındaki birçok çalışma ile paralellik göstermektedir (Michaelides, 2019; Wang ve diğ., 2015; Yang ve diğ., 2012). Araştırmalar özellikle küçük yaş gruplarında madde ifade etkisinin daha etkili olabileceğini belirtmektedir (Marsh, 1996; Michaelides, 2019; Weems ve diğ., 2003). Bunun nedeni, özellikle yaşı küçük bireylerin olumsuz yönde ifade edilmiş maddeleri anlamlandırmada daha fazla zorluk yaşaması olarak gösterilmektedir.

Çalışmada Model 5’in sonuçlarına göre öğrencilerin erken okuryazarlık aktiviteleri değişkeniyle SCM ve SCS’de bulunan madde ifade etkisi arasında manidar ve pozitif ilişki bulunmaktadır. Bu ilişkinin etki büyüklüğü ise düşük düzeydedir. Buna göre bu öğrenciler olumsuz yönde ifade edilmiş maddeleri yanıtlarken daha yüksek kategorileri tercih etmektedirler. Bunun yanında, çalışmadaki diğer değişkenler ile madde ifade etkisi arasında manidar ilişki bulunmamıştır. Tüm bu bulguların nedeni olarak erken okuryazarlıkla ilgili becerilerin okuduğunu anlamaya etkisinin zamanla azalıyor olması belirtilebilir. Bu konuda yapılan boylamsal çalışmalar bu durumu desteklemektedir (McTigue ve diğ., 2020; Roth ve diğ., 2002).

Bu çalışmanın bulguları, sınırlıklar çerçevesinde değerlendirilmelidir. Öncelikle, çalışmaya erken okuryazarlıkla ilgili sınırlı sayıda değişken dahil edilmiştir. Aynı zamanda dahil edilen değişkenler, öğrenciler ya da öğretmenler yerine ebeveynler tarafından yanıtlandırılan ölçme aracından elde edilmiştir. Alan yazındaki çalışmalar, ebeveynlerin ölçeklere verdikleri yanıtların sosyal beğenirlikten

etkilenileceğini göstermektedir (Huang, 2017). Son olarak, bu çalışmada öğrencilerin okuduğunu anlama becerilerine ilişkin başarı puanları bulunmamaktaydı. Bu nedenle, erken okuryazarlık becerileri yüksek düzeyde olan öğrencilerin okuduğunu anlamada ne kadar başarılı olduğu bilinmemekteydi.

Çalışmanın sınırlılıklarına rağmen araştırmacılara ve uygulayıcılara bazı öneriler sunulabilir. Birinci olarak, ölçek uyarlama ya da geliştirme çalışmalarında, özellikle uygulama yapılacak yaş grubu dikkate alınarak olumsuz yönde ifade edilen maddelerin incelenmesi önerilir. Ayrıca bu tür maddelerin öngörülen faktör yapısını tehdit edip etmediği de araştırılmalıdır. Eğer araştırmacılar ya da uygulayıcılar hali-hazırda kullanılan ve olumsuz yönde ifade edilmiş maddeler içeren ölçekleri kullanacaklarsa, bu ölçeklerde madde ifade etkisinin varlığını kontrol etmeleri uygun olacaktır. Eğer ölçeklerde madde ifade etkisi varsa alan yazında önerilen yöntemlerle madde ifade etkisi kaldırılarak yanıtlayıcıların ölçek puanları bu şekilde hesaplanmalıdır.

Comparison of Testlet Effect on Parameter Estimates Using Different Item Response Theory Models

Esin YILMAZ KOĞAR *

Abstract

In this study, the testlet effect was calculated for each testlet in the PISA 2018 reading literacy test, and it was examined whether this effect caused a difference in item and ability parameters. The data set was analyzed with a two-parameter logistic item response theory model and a two-parameter logistic testlet model. The results show that variances of testlet effects range from .100 to .432. When the item and ability parameter estimation results of the models were compared, it was determined that the item and ability parameters estimated from the two approaches were highly correlated with each other. It can be said that the item slope and item intercept parameters estimated from different models remained unaffected. However, when the local dependency assumption is not met, it was observed that the standard error values of the two-parameter model for the ability parameter were underestimated. The implications for the analysis and evaluation of the tests based on testlet are discussed. In conclusion, in this study, it was concluded that the testlet effect caused a difference in parameter estimates, but the local dependence among the items was negligible because of the small testlet effects.

Key Words: Local item dependency, item response theory, testlet response theory, testlet effects, PISA.

INTRODUCTION

A testlet is defined as a cluster of items that share a common stimulus (Wainer & Kiely, 1987). This common stimulus can be presented as a passage, scenario, table, or figure. Testlets are widely used in testing for several reasons such as ensuring the effective use of the time required for the test application, reducing the context effect that may arise from the content of the items in the test, eliminating the concerns that a single independent item may be too atomistic (measuring a very specific or narrow concept) because of its nature (Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007). However, if different items are collected in the same testlet, these items may be related to each other beyond the effect of the latent trait that is tried to be measured. This situation, known as local item dependency (LID), leads to the violation of the local independence assumption of standard item response theory (IRT) models. For example, the performance of students in a reading comprehension test may be affected by their interest in or knowledge of reading passages, as well as their reading skills (Yen, 1993). Therefore, items in the same set of items may be locally dependent.

The local item dependency (LID) between testlet items is called the testlet effect (Wainer & Kiely, 1987). Bradlow, Wainer, and Wang (1999) proposed a new model by adding this effect as a parameter to the 2-parameter logistic model (Birnbaum, 1968, 2PLM). In this model, which is called the testlet response theory (TRT) model, there is a random-effects parameter, γ , that considers account the dependencies between the items in the same testlet. In the standard 2PL IRT model, there are item difficulty and item discrimination parameters, and it is assumed that there is no local dependence between items. In the TRT model, calculations are made by including item difficulty and item discrimination parameters, as well as a random effect parameter. The 2PL TRT model, which is developed in the standard 2PL IRT model, can be written as (Li, Bolt, & Fu, 2006; Ip, 2010);

* Asst. Prof. Dr., Niğde Ömer Halisdemir University, Faculty of Education, Niğde-Turkey, esinyilmazz@gmail.com, ORCID ID: 0000-0001-6755-9018

To cite this article:

Yılmaz Koğar, E. (2021). Assessing testlet effect on parameter estimates obtained from item response theory models. *Journal of Measurement and Evaluation in Education and Psychology*, 12(3), 254-266. doi: 10.21031/epod.948227

Received: 5.06.2021

Accepted: 21.07.2021

$$P(Y_{ij} = 1 | \theta_j, \gamma_{jd(i)}) = \frac{\exp(a_i (\theta_j - b_i - \gamma_{jd(i)}))}{1 + \exp(a_i (\theta_j - b_i - \gamma_{jd(i)}))} \quad (1)$$

where $P(Y_{ij} = 1)$ is the probability that examinee j answers item i correctly, θ_j is the ability of examinee j , a_i denotes the discrimination parameter of item i , b_i is the difficulty of item i . The testlet effect $\gamma_{jd(i)}$ for examinee j is such that his or her response to item i is nested within testlet $d(i)$, and this testlet effect is assumed to be independent of the latent trait θ .

It has been thought that the use of standard IRT models for these tests may be insufficient since the LID assumption has been violated in the tests involving testlets. Therefore TRT models have become a frequently used model in research to testlet effect (DeMars, 2006; Eckes, 2014; Geramipour, 2021; Min & He, 2014; Özdemir, 2017; Paap & Veldkamp, 2012; Wainer & Wang, 2000; Yılmaz Kogar & Kelecioğlu, 2017). Glas, Wainer, and Bradlow (2000) examined in their simulation study that when the testlet effect was ignored and the standard IRT model was used, the mean absolute errors of discrimination and difficulty parameter estimation were poorly predicted. Wainer and Wang (2000), in their study based on TOEFL results, determined that the testlet model developed by adding the γ parameter, expressed as the random testlet effect, to the standard 3PL IRT model, gave better results in parameter estimation. Özdemir (2017) conducted a study in which he analyzed the English Proficiency Test data with the TRT model, the dichotomous and polytomous IRT models. In this study, he compared item and ability parameter estimations and determined that the results differed, especially for item parameters. Studies in the literature show that the use of standard IRT models when LID is present can lead to problems such as biased item parameter estimates, overestimation of the accuracy of ability estimates, overestimation of test reliability and test information, and underestimation of standard errors for ability parameter (Sireci, Thissen, & Wainer, 1991; Wainer et al., 2007; Yen & Fitzpatrick, 2006). Based on the results of these studies, it can be said that serious problems may be encountered for the psychometric properties of the tests when LID is ignored. This may lead to incorrect results regarding the interpretation and use of test scores.

Testlets, which are based on a common stimulus and group of items, are used in many large-scale tests because of the previously specified advantages. One of these tests is the PISA (Program for International Student Assessment) applied on the international platform by OECD (Organisation for Economic Co-operation and Development). This application, which evaluates the knowledge and skills of 15-year-old students every three years, focused on reading literacy skills in 2018. Testlets are used in tests that measure language skills, such as reading comprehension. However, in such items, some students have a special interest or better prior background knowledge in a passage than other students, in this situation, they are likely to perform better on the items related to this passage than on other items of the same difficulty level, or they tend to perform better than other students with the same general ability level (Li, 2017, p.1). Therefore, testlets lead to the emergence of additional variance sources, such as content knowledge in an item response function (Chen & Thissen, 1997). However, it is still not commonly enough to perform analyzes through the models that take this effect. The current study is aimed to fill this gap.

PISA applications, which are very important to national and international platforms, are classified as low-stake tests because the important personal decisions associated with the test performance of the participants are not taken. However, the role of these applications in the educational policies of countries is great. IRT approach is used for item and ability estimates in PISA; these models are not special IRT models developed for testlets. In this respect, it is a condition that the results obtained from the standard IRT models and the results obtained from TRT models will change all interpretations. Because it is desirable to be estimated by the least amount of error to achieve a high degree of accuracy. If the LID is a large effect on the estimates of the testlets, this may be compromised.

This study is aimed to calculate the LID magnitude caused by testlets and to compare the effect of this magnitude on parameter estimates and test precision. The following research questions have been established to address these situations:

1. What is the LID level of testlets included in the PISA 2018 reading literacy test?
2. Do the person and item parameters obtained with the standard IRT model and TRT model differ?

By determining the level of testlet variances obtained through the real data sets with these research questions, it is aimed to make an inference about the situations in which the use of TRT models proposed in the literature may be necessary. Also, this study aims to help researchers, especially those used to standard IRT models, to better understand and interpret testlet models because TRT models are less known and less used models than standard IRT models.

METHOD

Participants

PISA application is carried out on 15-year-old students enrolled in formal education. Schools and students participating in the PISA research are determined by the OECD randomly. There are more than 600,000 students from 79 countries and economies participating in the PISA 2018 application (Organisation for Economic Co-operation and Development, 2019). In this study, countries participating in PISA 2018 application as computer-based administrations were preferred. The data of these countries were examined in terms of the same test design and testlets, and analyses were carried out on 3105 students, who were suitable for the study.

Data Sources and Measures

In PISA 2018 application, the main domain is reading literacy. In PISA 2018, a multistage adaptive test (MSAT) design was used to measure reading skills. The MSAT design for the PISA 2018 main survey consisted of three stages (Core stage, Stage 1 and, Stage 2) and 245 items. Different designs were created by applying these stages in different orders. In this design, between 33 and 40 items were applied to each student, depending on which test was taken at each stage. The data used in this study were obtained from design A (Core> Stage 1> Stage 2) applied to 75% of the students. From 64 different ways defined for design A, the selected path is RC1 for the core stage, R15H for stage 1, and R21H for stage 2. For detailed information, it is recommended to consult the report of Yamamoto, Shin, and Khorramdel (2019).

The items in the reading literacy test are in a format that includes constructed response or selected response. However, this study focused on only multiple-choice and dichotomous items because the models used in the study were developed for the items scored dichotomously. The data in PISA applications are open to everyone's use. However, the items are not shared because the items in cognitive instruments are used in other years. For this reason, only data coding was considered for the testlet decision regarding the items. The "label" section of the reading literacy test has been examined in the SPSS format and assumed that the items in the same label are testlets. After this review, 39 items comprising seven testlets were used in the study. The reason for the use of PISA data in the study is that it provides a real set of data in testing and applies to many people.

The data of the study were accessed at <https://www.oecd.org/pisa/data/>

Data Analysis

Two different measurement models were used in the study: (a) standard 2PL IRT model, (b) 2PL testlet response model. The reason 2PL models are used in the study is that when 3PL is used in TRT models, convergence problems can be experienced for parameter estimation (Eckes, 2014). In this study, the item and ability parameters estimate obtained from the standard IRT model and TRT model was compared with the corresponding standard errors. Root Mean Square Error (RMSE) was examined to

compare the capability parameters estimated from different IRT models. RMSE values are calculated by taking the square root of the mean square of the standard errors of the ability parameters. Besides, to better understand the degree of agreement between the estimates, correlations related to the estimates of the two models were calculated, and statistics based on mean differences were used (Mean Difference-MD, Mean Absolute Difference-MAD, Root-Mean-Square Difference-RMSD).

Analyses were performed using the mirt package (Chalmers, Pritikin, Robitzsch, & Zoltak, 2015) included in the R software. mirt is a package developed for multidimensional IRT models. Therefore, it includes slope and intercept parameters as item parameters. For the unidimensional 2PL model, the slope parameter is the same as the discrimination parameter (a_i), while the intercept parameter (d_i) is calculated over the discrimination and difficulty parameter (b_i) ($d_i = -a_i b_i$). In this study, the intercept parameter transformation is used instead of the difficulty parameter. The item intercept parameter is interpreted as item easiness and is the opposite of the item difficulty parameter. In general, a high value means that the item is easy (Reckase, 2009). The item slope parameter is interpreted as the item discrimination parameter. Higher values indicate that the item is more distinctive (Baker, 2001).

It is also assumed that the population ability distribution in the pack follows a normal distribution. Therefore, there is a normal distribution with mean and standard deviation equal to 0 and 1, respectively, for model identification purposes in IRT calibrations (Paek & Cole, 2020). In this way, parameter estimates obtained from different IRT models are provided to be on the same scale (Li, Li, & Wang, 2010). Also, the calculation of IRT scale scores was performed using the EAP (expected a posteriori) method.

RESULTS

The current study, firstly, analysis results based on the TRT model are presented and focus on testlet effect variance as an indicator of LID for each testlet. Then, the item parameter estimates obtained from the TRT model and the standard IRT model were compared, and the RMSE values showing the precision of these estimates were calculated for each model. Then, various statistics based on correlation values and mean differences are given to examine the fit between models. The same operations were done for the estimations regarding the ability parameter.

The Testlet Effects

The testlet effect variance shows the degree of local dependency among items included in a particular testlet. When the testlet effect variance is zero, there is no local dependence between items. The more this variance exceeds zero, the higher the degree of LID. However, there are different approaches to interpret this value. In simulation studies, it is generally stated that variances below .25 can be considered negligibly small (Glas et al., 2000; Wang & Wilson, 2005). For the testlet effect variance, values of .50 and above are considered to be more important (Wang & Wilson, 2005; Wainer et al., 2007). Table 1 shows the magnitudes of γ and standard errors of testlet effects.

Table 1. Testlet Statistics

Testlet	Number of Items	Testlet Variance	Standard Error
Testlet 1	4	.173	.099
Testlet 2	5	.432	.142
Testlet 3	6	.088	.077
Testlet 4	3	.157	.123
Testlet 5	2	.200	.235
Testlet 6	7	.100	.044
Testlet 7	6	.365	.070

As shown in Table 1, some testlets have much higher LID than others. The variance of the testlet effect for testlet 2 (the code of the testlet is “South Pole”) is .489, which is much greater than for other testlets. However, it is seen that all testlet effect variances are less than .50. Looking at the estimations for standard errors, it can be said that these values are not very high, and therefore each testlet effect variance is estimated precisely.

Item Parameter Estimates

The standard IRT model which ignores LID and TRT model item parameters and RMSE values are showed in Table 2.

Table 2. Summary Statistics for Estimated Item Parameters

Model	Slope					Intercept				
	Mean	SD	Min	Max	RMSE	Mean	SD	Min	Max	Mean
IRT	.87	.31	.37	1.57	.08	1.27	1.49	-1.83	5.30	.09
TRT	.87	.33	.35	1.71	.09	1.32	1.58	-1.88	6.01	.13

Note: SD = standard deviation, Min = minimum, Max = maximum, RMSE = Root Mean Square Error.

The summary statistics are shown in Table 2 show to a very high correspondence between the item parameters estimated by the standard IRT and TRT models. Especially item slope parameters were estimated with extreme precision by both models but item intercept parameters, the precision was somewhat lower but still very high. Besides, when the RMSE values are examined, it is seen that the values obtained from the TRT model are higher.

Correlation values and mean differences calculated to determine the amount of agreement of item parameters obtained from different models are given in Table 3.

Table 3. Correlations and Mean Differences for Item Parameter Estimates from Different Models

Parameter	Correlation	MD	MAD	RMSD
Slope	.996	-.009	.032	.034
Intercept	.998	-.095	.103	.303

Note: MD = mean differences, MAD = mean absolute differences, RMSD = root mean square differences.

Table 3 presents the correlations and difference-based statistics for item slope and intercept estimates, respectively. When the correlation values in this table are examined, it is seen that the item parameters obtained from both models are highly correlated. Mean differences between the item parameters obtained from the two models were also calculated to see if one model produced higher or lower parameters than the other model. It can be seen that the average differences for both parameters are very small. However, when looking at the RMSD values, it can be said that the item parameters are affected by the testlet structure. It is seen that testlet structure in the test can produce biased results especially for the intercept parameter.

The relationships of the estimations on item parameters obtained from the IRT model and TRT model are shown in Figure 1.

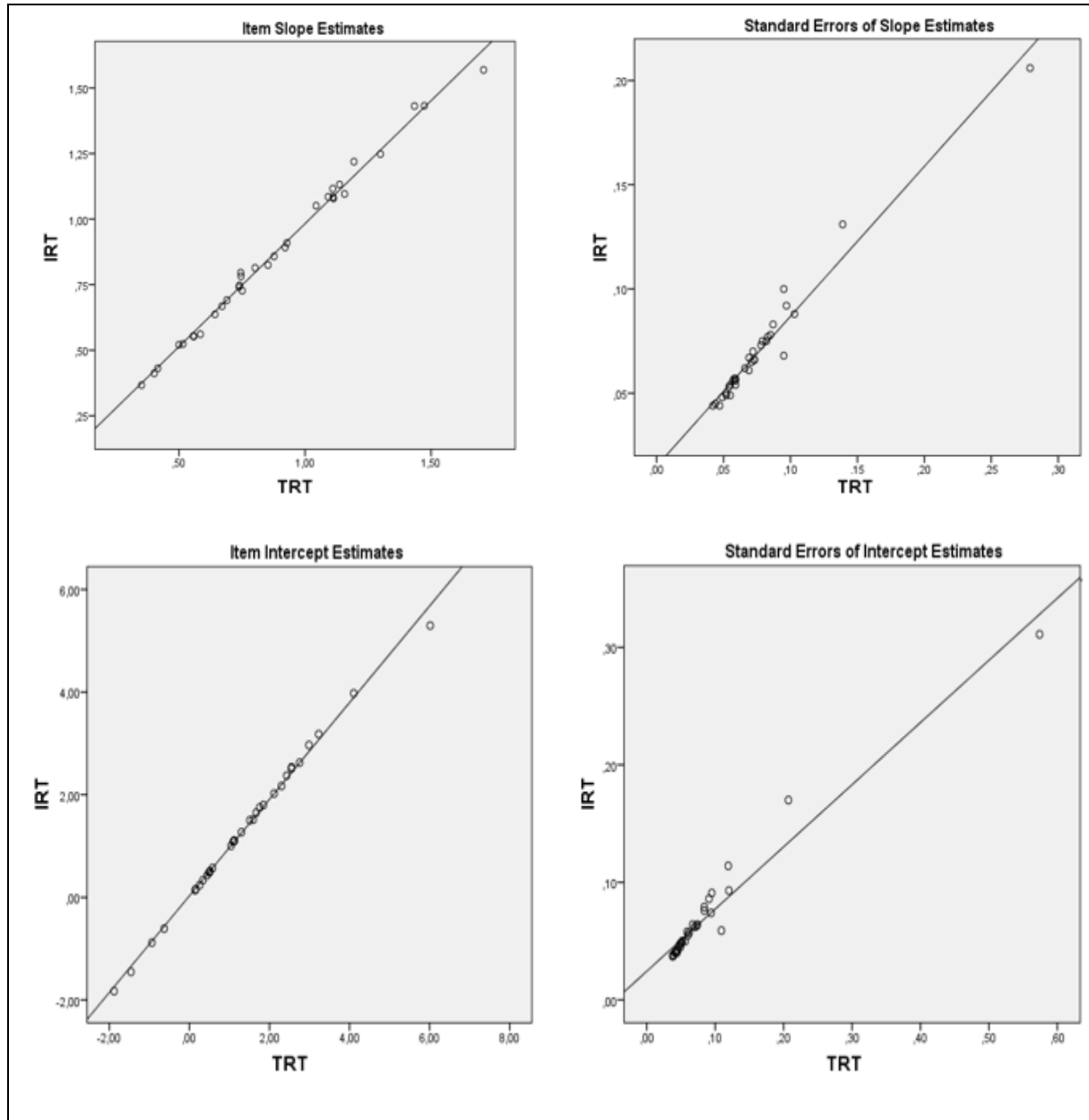


Figure 1. Item Slope and Item Intercept Estimates Under the Standard 2PL Model and Testlet Response Model

When Figure 1 is analyzed, it can be said that item parameter estimates are similar in both models. However, while the standard errors related to the item slope parameters are still similar, there is a slight difference in the standard errors for the item intercept parameter. The standard errors estimated from the standard IRT model for item slope parameters vary between .04 and .21, while the standard errors estimated from the TRT model vary between .04 and .28. The standard errors estimated from the standard IRT model for item intercept parameters vary between .04 and .31, while the standard errors estimated from the TRT model vary between .04 and .57. Therefore, it can be said that the standard IRT model underestimated the measurement error.

Person Ability Estimates

Descriptive statistics for the ability parameters obtained from two IRT models and the RMSE values for the accuracy of this estimate are given in Table 4.

Table 4. Summary Statistics for Person Ability Estimates

Model	Minimum	Maximum	SD	RMSE
IRT	-3.04	2.29	.87	.49
TRT	-2.70	2.15	.81	.58

Note: The mean of the ability distribution was fixed at 0 for estimation purposes for the two models, SD = standard deviation; RMSE = root mean square error.

When looking at the minimum and maximum values and standard deviation values for the ability estimation in Table 4, the estimates from the 2PL IRT model showed a somewhat larger variation than the estimates from the testlet model. When the RMSE values are examined, the higher measurement precision was obtained from the 2PL IRT model compared to the TRT model. In addition to these values, correlation and mean differences were calculated to show the fit between the ability parameters estimated from the two models. It was determined that there is a high correlation between ability parameters obtained from independent items and the TRT model ($r = .996$). The value found for MAD is .098, and the value found for RMSD is .123. For this reason, it can be said that the ability parameters estimated from both models are similar. Figure 2 shows the scatter plots of ability estimates obtained from both models and the standard errors of these estimates.

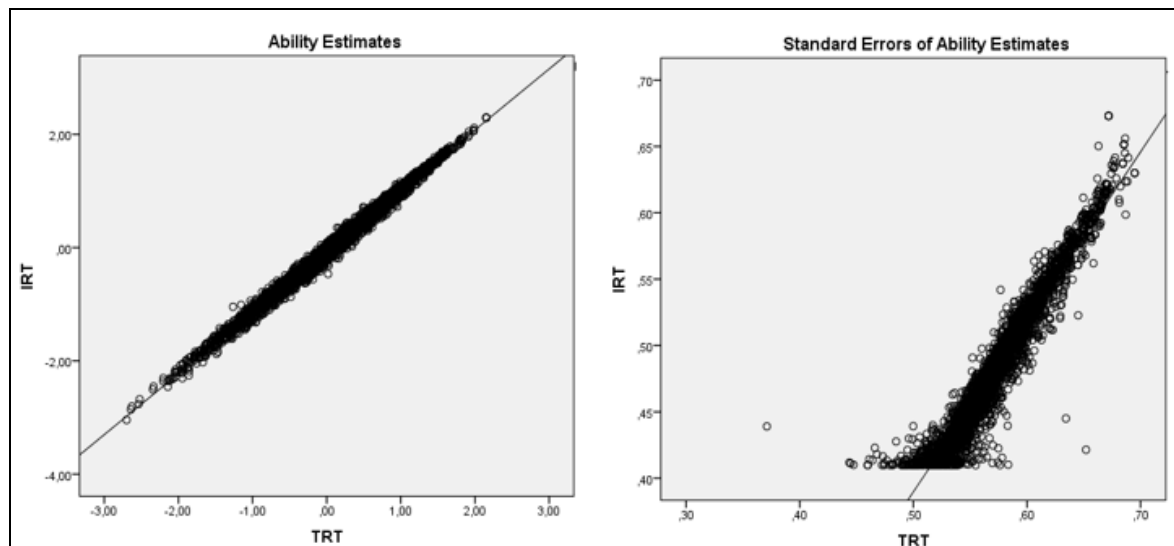


Figure 2. Person Ability Estimates and Associated Standard Errors under the Standard IRT Model and the Testlet Response Model

On the left of Figure 2, the distribution of ability estimates of different models and on the right side, the distribution graphs of the standard errors of the relevant parameter are shown. It can be said that the estimates of the two models are almost the same according to the scatter plot of the ability parameters obtained from the standard IRT model and TRT model. However, when the graph regarding the standard errors is examined, it is seen that the standard IRT model estimates the errors less. While the standard errors estimated from the IRT model ranged from .41 to .67, the standard errors estimated from the TRT model ranged from .37 to .69. Therefore, it can be said that the standard IRT model underestimated the measurement error.

DISCUSSION and CONCLUSION

The aim of this study is to calculate LID magnitude resulting from the testlets in the PISA 2018 reading literacy test and to compare the effect of this size on parameter estimates and test accuracy. For this

purpose, item and ability parameter estimations were performed using the IRT model with local independency assumption and the TRT model.

First, the LID status among the items was examined by calculating the testlet effect variance. It was determined that the testlet effects found for the seven testlets were lower than .50. Therefore, it can be said that there is no strong testlet effect in the data set. In studies conducted on real data in the literature, it has been observed that testlet effect variances are lower than .50 (Baghaei & Ravand, 2016, Chang & Wang, 2020; Eckes, 2014).

Then, the item parameters estimated on the standard IRT model and TRT models were compared. The results obtained show that the item parameter estimates are similar. In general, the RMSDs between the item parameters estimated from the two models were low. It was also determined that the slope parameter gives more similar results than the intercept parameter. However, this result differs from the results of the study conducted by Min and He (2014). Comparing the item parameters of different IRT models, the researchers stated that the slope parameter was estimated more suspiciously than the intercept parameter. However, in this study, the bifactor model, another model used in testlets, was chosen as the basic model, and this model was compared with other models. In the present study, the bifactor model was excluded. The difference observed may be due to comparison with different models.

Correlations between item parameter estimates obtained from both models are quite high. DeMars (2006), in his research with PISA 2000 data, used both mathematics and reading literacy data to examine the ability estimations of the independent item model and testlet effect model and stated that the correlations between these estimations were close to 1. A similar result has been observed in other studies (Baghaei & Ravand, 2016; Eckes, 2014; Eckes & Baghaei, 2015; Yılmaz Kogar & Kelecioğlu, 2017).

For the last stage of the research, the estimates regarding the ability parameters were examined. Although the ability parameter results obtained from the standard IRT and TRT models are similar, it is seen that the results of the standard IRT model differ more. However, considering the correlation for this parameter and the statistics based on the mean differences of these estimates, it was determined that the IRT and TRT models show high correlation and are quite compatible with each other with small RMSD values. This finding is in line with the findings of the studies conducted by Eckes (2014) and Özdemir (2017). Besides, standard errors related to the ability parameter are estimated higher in the TRT model. In the literature, it is stated that if the item team effect is ignored, the standard error for the ability parameter is underestimated (Chang & Wang, 2010; Wainer, Bradlow, & Du, 2000; Wainer & Wang, 2000).

Conclusion and Suggestions

Testlets allow more than one item to be asked based on the same stimulus, allowing more than one information to be collected from a stimulus, thus improving the efficiency of the test (information per unit time) (Wainer et al., 2000). Therefore, the use of such items in tests is inevitable. However, it is also necessary to deal with the violation of the local independence assumption of testlet items. To this end, it is important to determine in which cases breaking this assumption will affect the results.

The current study was determined that the results obtained from the standard IRT model and the TRT model are quite close to each other. This result is similar to the studies conducted on the real data set (Baghaei & Ravand, 2016; Demars, 2006; Eckes, 2014; Eckes & Baghaei, 2015; Özdemir, 2017; Yılmaz Kogar & Kelecioğlu, 2017). The reason why the result is this way is probably the small variance of the testlet in the data set used in this study because Glas et al. (2000) stated that the testlet effect variances lower than .50 had a negligible effect on the results. They also stated that in this case, standard IRT models, such as 2PL or 3PL could be used without compromising the quality of the parameter estimates. However, even in studies with a high testlet effect, correlations between standard IRT models and TRT models were high (Baghaei & Ravand, 2016; Özdemir, 2017). Beside, it was observed that there were partial variations in RMSE and standard errors obtained from the parameters. According to DeMars (2006), although the complex model results in slightly higher RMSE than the less complex model, this

is not a bias. Differences in standard errors were observed, especially in the ability parameter. Such differences can lead to negative consequences when it comes to high-risk decisions (Baghaei & Ravand, 2016). Besides, this can cause serious problems when using computer adaptive tests, which are test termination criteria, the standard error of ability estimates.

As a result, when there is a very strong dependency between the items in the tests, standard IRT models will not give appropriate results for testlet as they neglect this addiction because the studies conducted show that neglecting the assumption of local independence violation causes overestimation of reliability or knowledge and underestimation of standard error of ability estimation (Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen, 1993). However, researchers who have difficulty using more complex models when the testlet effect is low can use standard IRT models since when the testlet effect is low, it can be said that these models do not make very different predictions from the TRT models. Researchers working with testlets are primarily recommended to examine the testlet variance. Then, if the testlet effect is low, it can be said that standard IRT models can be used for parameter estimates. If there is a high testlet effect, TRT models are required.

Limitations

Despite the contribution of this research to the field, it has several limitations that require further research. Since real data was used in the study, the results of the current situation were examined and the testlet effect variance was estimated to be low. With different studies it can be examined how high these effects can be based on real data. Also, instead of determining only this effect, studies can be conducted to determine the source of the variance created by this effect. For this purpose, the characteristic features of the testlet can be examined using real data, where each item in the test can be accessed. However, since not all the items could be accessed in PISA applications, the characteristics of the testlet items could not be examined in this study. Also, only dichotomous items were used in the study. In future research, the regulations that will consider account the polytomous items can be made.

In the current study, the 2PL TRT model, one model dealing with testlets, was used. TRT models are a limited form of bifactor models. For this reason, the testlet effect can also be handled with bifactor models. In the future, similar studies can be done using the bifactor model and models containing more parameters.

REFERENCES

- Baghaei, P. & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica*, 37(1), 85-104.
- Baker, F. B. (2001). *The basics of item response theory*. College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.
- Chang, Y., & Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment*. Paper presented at 4th IEA International Research Conference, Gothenburg, Sweden. Retrieved from http://www.iea-irc.org/fileadmin/IRC_2010_papers/PIRLS/Chang_Wang.pdf
- Chalmers, P., Pritikin, J., Robitzsch, A., & Zoltak, M. (2015). Package 'mirt'. Retrieved January 10, 2021, from <https://mran.microsoft.com/snapshot/2014-12-27/web/packages/mirt/mirt.pdf>
- Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. <https://doi.org/10.3102/10769986022003265>
- DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43(2), 145-168.
- Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, 31(1), 39-61.
- Eckes, T. & Baghaei, P. (2015). Using testlet response theory to examine local dependence in C-tests. *Applied Measurement in Education*, 28(2), 85-98.

- Geramipour, M. (2021). Rasch testlet model and bifactor analysis: how do they assess the dimensionality of large-scale Iranian EFL reading comprehension tests?. *Language Testing in Asia*, 11(1), 1-23. <https://doi.org/10.1186/s40468-021-00118-5>
- Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271-287). Boston, MA: Kluwer-Nijhoff.
- Ip, E. H. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement*, 34(7), 467-482.
- Li, F. (2017). An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory. *ETS Research Report Series*, (1), 1-25. <https://doi.org/10.1002/ets2.12151>
- Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30(1), 3-21.
- Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (ETS RR-10-21). Princeton, NJ: Educational Testing Service.
- Min, S. & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, 31(4), 453-477.
- Organisation for Economic Co-operation and Development (2019). *PISA 2018 assessment and analytical framework*. Paris: OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- Özdemir, B. (2017). Examining testlet effects in english proficiency test: A Bayesian testlet response theory approach. In I. Koleva & G. Duman (Eds.), *Educational Research and Practice*, (pp. 425-437). Sofia: ST. Kliment Ohridski University Press.
- Paap, M. C., & Veldkamp, B. P. (2012). *Minimizing the testlet effect: Identifying critical testlet features by means of tree-based regression*. Psychometrics in Practice at RCEC, 63. Retrieved January 12, 2021, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1001.1923&rep=rep1&type=pdf#page=71>
- Paek, I., & Cole, K. (2019). *Using R for item response theory model applications*. London: Routledge.
- Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). New York, NY: Springer.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237-247.
- Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory. An analog for the 3PL useful in testlet-based adaptive testing. In W. J. van der Linden & G. A. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 245-269). Springer, Dordrecht. https://doi.org/10.1007/0-306-47531-6_13
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201.
- Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37(3), 203-220.
- Wang, W. C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, 29(4), 296-318.
- Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice*, 37(4), 16-27.
- Yen, W. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.
- Yılmaz Kogar, E., & Kelecioğlu, H. (2017). Examination of different item response theory models on tests composed of testlets. *Journal of Education and Learning*, 6(4), 113-126.

Parametre Tahminleri Üzerindeki Madde Takımı Etkisinin Farklı Madde Tepki Kuramı Modelleri Kullanılarak Karşılaştırılması

Giriş

Madde takımı (testlet), ortak bir uyararı paylaşan maddeler kümesi olarak tanımlanır (Wainer ve Kiely, 1987). Bu ortak uyararı bir metin, senaryo, tablo ya da şekil olarak sunulabilir. Madde takımları, test

uygulanması için gerekli zamanın etkili kullanılmasını sağlaması, testteki maddelerin içeriğinden kaynaklı oluşabilecek içerik etkisini azaltması, tek bir bağımsız maddenin doğası gereği fazla atomistik (çok özel veya dar bir kavramı ölçme) olabileceğine dair endişeleri ortadan kaldırması gibi çeşitli nedenlerle testlerde oldukça kullanılmaktadır (Wainer, Bradlow ve Du, 2000; Wainer, Bradlow ve Wang, 2007). Ancak farklı maddelerin aynı madde takımında toplanması durumunda bu maddeler, ölçülmeye çalışılan gizil özelliğin etkisinin ötesinde birbirleriyle ilişkili olabilir. Yerel madde bağımlılığı olarak bilinen bu durum standart madde tepki kuramı (MTK) modellerinin yerel bağımsızlık varsayımının ihlâl edilmesine yol açar. Örneğin okuduğunu anlama becerisinin ölçüldüğü bir teste yer alan maddelerde öğrencilerin performansı, okuma becerisinin yanı sıra okuma parçası içeriğine olan ilgisinden veya bilgisinden etkilenebilir (Yen, 1993). Bu nedenle de aynı madde takımında yer alan maddeler yerel bağımlı olabilir.

Madde takımlarından kaynaklanan yerel madde bağımlılığına madde takımı etkisi denir (Wainer ve Kiely, 1987). Bradlow, Wainer ve Wang (1999), 2 parametrelili lojistik modele (Birnbaum, 1968, 2PLM) bu etkiyi de bir parametre olarak eklemiş ve yeni bir model önermişlerdir. Madde takımı tepki kuramı (MTTK) olarak isimlendirilen bu modelde, aynı madde takımında yer alan maddeler arasındaki bağımlılıkları da hesaba katan bir rastgele etkiler parametresi, γ , bulunur. Standart 2PL MTK modelinde madde güçlük ve madde ayırt edicilik parametreleri bulunmakta ve maddeler arasında yerel bağımlılık olmadığı varsayılmaktadır. MTTK modelinde ise madde güçlük ve madde ayırt edicilik parametrelerinin yanı sıra bir rastgele etki parametresi de dâhil edilerek hesaplamalar yapılır.

Madde takımı etkisini göz önüne alan MTTK modelleri araştırmalarda sıklıkla kullanılan bir model hâline gelmiştir (DeMars, 2006; Eckes, 2014; Min ve He, 2014; Paap ve Veldkamp, 2012; Wainer ve Wang, 2000). Glas, Wainer ve Bradlow (2000) yaptıkları simülasyon çalışmasında, madde takımı etkisinin görmezden gelindiği ve standart MTK modelinin kullanıldığı durumda, ayırt edicilik ve güçlük parametre kestirimlerinin ortalama mutlak hatasının kötü tahmin edildiğini belirlemişlerdir. Wainer ve Wang (2000) TOEFL sonuçları üzerinden yürüttükleri çalışmada Standart 3PL MTK modeline tesadüfi madde takımı etkisi olarak ifade edilen γ parametresinin eklenmesiyle geliştirilen madde takımı modelinin parametre kestirimlerinde daha iyi sonuç verdiğini belirlemişlerdir. Alanyazında yer alan araştırmalar, yerel madde bağımlılığı mevcutken standart MTK modellerinin kullanılmasının yanı sıra madde parametre kestirimlerine, yetenek kestirimlerinin kesinliğinin fazla tahmin edilmesine, test güvenilirliğinin ve test bilgilerinin fazla tahmin edilmesi gibi sorunlara yol açabildiğini göstermektedir (Sireci, Thissen ve Wainer, 1991; Wainer vd., 2007; Yen ve Fitzpatrick, 2006). Bu araştırmaların sonuçlarına dayanarak yerel madde bağımlılığı göz ardı edildiğinde testlerin psikometrik özellikleri için ciddi sorunlarla karşılaşılacağı söylenebilir. Bu durum ise test puanlarının yorumlanması ve kullanılmasıyla ilgili yanlış sonuçlar doğurabilir.

Birçok geniş ölçekli testte, ortak bir uyarana dayanan ve madde takımı olarak adlandırılan madde grupları kullanılmaktadır. Özellikle okuduğunu anlama becerileri için geliştirilen testlerde madde takımlarına oldukça yer verilir. Ancak bu madde takımlarının neden olduğu madde takımı etkisi, bir madde cevap fonksiyonunda ek bir varyans kaynağı oluşturur. Buna karşın bu etkiyi göz önüne alan modeller üzerinden analizler gerçekleştirmek hâlâ yeterince yaygın değildir. Bu çalışma ile bu boşluğun doldurulmasına katkı sağlamak hedeflenmektedir. Bu çalışmada; madde takımlarından kaynaklı oluşan yerel madde bağımlılığı büyüklüğünü hesaplamak, bu büyüklüğün parametre tahminleri ve test kesinliği üzerindeki etkisini karşılaştırmak amaçlanmaktadır. Bu durumları ele almak için aşağıdaki araştırma soruları oluşturulmuştur:

1. PISA 2018 okuma becerileri testinde yer alan madde takımlarının yerel madde bağımlılığı derecesi nedir?
2. Standart 2-PL MTK modeliyle elde edilen kişi ve madde parametreleri ile 2-PL MTTK modeliyle elde edilen kişi ve madde parametreleri farklılaşmakta mıdır?

Yöntem

PISA uygulaması, örgün öğretimde kayıtlı olan 15 yaş grubu öğrencilerin katıldığı bir uygulamadır. PISA araştırmasına katılacak okul ve öğrenciler, OECD tarafından seçkisiz yöntemle belirlenmektedir. PISA 2018 uygulamasına toplam 79 ülke ve ekonomiden katılan 600.000'den fazla öğrenci bulunmaktadır (Organisation for Economic Co-operation and Development, 2019). Bu çalışmada PISA 2018 uygulamasına bilgisayar tabanlı değerlendirme şeklinde katılan ülkeler tercih edilmiştir. Bu ülkelerin verileri test düzeninin ve madde takımlarının aynı olması bakımından incelenmiş ve araştırmanın amacına uygun olan 3105 öğrenci üzerinden analizler gerçekleştirilmiştir.

PISA 2018 uygulamasında ağırlıklı alan okuma becerileridir (reading literacy). PISA 2018'de okuma becerilerini ölçebilmek için çok aşamalı uyarlanmış test (multistage adaptive test-MSAT) deseni kullanılmıştır. Bu deseni içeren uygulamada okuma becerileri alanı için toplam 245 madde bulunmaktadır. Maddeler; temel, 1. aşama ve 2. aşama olacak şekilde üç aşamada yer alacak şekilde yapılandırılmıştır. Bu aşamaların farklı sıralarda uygulanmasıyla farklı düzenler oluşturulmuştur. Bu desende her öğrenciye her aşamada hangi testin alındığına bağlı olarak 33 ile 40 arasında madde uygulanmıştır. Bu çalışmada kullanılan veriler, öğrencilerin %75'ine uygulanan A düzeninden (Core>Stage 1>Stage 2) elde edilmiştir. A düzeni için tanımlanan 64 farklı yoldan ise seçilen yol temel aşama için RC1, 1. aşama için R15H ve 2. aşama için R21H şeklindedir. Ayrıntılı bilgi için Yamamoto, Shin ve Khorramdel'in (2019) raporuna bakılması önerilir.

Okuma becerileri testinde yer alan maddeler seçme gerektiren ya da öğrencinin cevabı kendisinin yapılandırmasını gerektiren formattadır. Ancak bu çalışmada yalnızca çoktan seçmeli ve ikili puanlanan maddeler üzerine odaklanılmıştır. Çalışmada farklı sayıda madde içeren 7 madde takımının oluşturduğu toplam 39 madde kullanılmıştır.

Çalışmada iki farklı ölçme modeli kullanılmıştır: (a) 2PL Madde takımı tepki modeli (Wainer et al.,2007), (b) standart 2PL MTK modeli (Birnbaum, 1968). Çalışmada 2PL modellerinin kullanılmasının nedeni, MTTK modellerinde 3PL kullanıldığında parametre kestirimleri için yakınsama problemi yaşanabilmesidir (Eckes, 2014). Bu çalışmada standart 2PL MTK ve 2PL MTTK modellerinden elde edilen madde ve yetenek parametreleri kestirimleri ile bunlara karşılık gelen standart hatalar karşılaştırılmıştır. Farklı MTK modellerinden kestirilen yetenek parametrelerini karşılaştırmak için hataların ortalama karekökü (RMSE) incelenmiştir. RMSE değerleri yetenek parametrelerinin standart hatalarının karesinin ortalamasının karekökü alınarak hesaplanmıştır. Ayrıca kestirimler arasındaki uyuma derecesini daha iyi anlamak için iki modelin kestirimlerine ilişkin korelasyonlar hesaplanmış ve ortalama farklılıklarına dayalı istatistikler kullanılmıştır (MD, MAD, RMSD). RMSD, iki modelden kestirilen parametrelerine ilişkin hatalar farkının karesinin ortalaması alınarak elde edilmiştir. Analizler R programında mirt paketi (Chalmers vd., 2015) üzerinden gerçekleştirilmiştir.

Sonuç ve Tartışma

Bu çalışmanın amacı; PISA 2018 okuma becerileri testindeki madde takımlarından kaynaklı oluşan yerel madde bağımlılığı büyüklüğünü hesaplamak, bu büyüklüğün parametre tahminleri ve test kesinliği üzerindeki etkisini karşılaştırmaktır. Bu amaçla madde ve yetenek parametresi kestirimleri yerel bağımsızlık varsayımı bulunan MTK modeli ile MTTK modeli kullanılarak gerçekleştirilmiştir.

İlk olarak madde takımı etki varyansı hesaplanarak maddeler arasındaki yerel madde bağımlılığı durumu incelenmiştir. Yedi madde takımı için bulunan madde takımı etkisi düşük düzeydedir. Bu nedenle veri setinde güçlü bir madde takımı etkisinin olmadığı söylenebilir. Literatürde gerçek veriler üzerinden yapılan çalışmalarda da madde takımı varyanslarının .50'den düşük olduğu gözlenmiştir (Baghaei ve Ravand, 2016, Chang ve Wang, 2020; Eckes, 2014).

Daha sonra MTK ve MTTK modeli üzerinden kestirilen madde parametreleri karşılaştırılmıştır. Elde edilen sonuçlar madde parametre kestirimlerinin benzer olduğunu göstermektedir. Genel olarak, iki modelden tahmin edilen madde parametreleri arasındaki RMSD'lerin küçük olduğu belirlenmiştir.

Ayrıca a parametresinin, d parametresine göre daha benzer sonuçlar verdiği belirlenmiştir. Her iki modelde elde edilen madde parametre kestirimleri arasındaki korelasyonlar ise oldukça yüksektir. DeMars (2006) PISA 2000 verisiyle yaptığı araştırmada hem matematik hem okuma verileri için MTTK modeli ile standart MTK'nin yetenek kestirimlerinin korelasyonlarının 1'e yakın olduğunu belirtmiştir.

Yetenek parametrelerine ilişkin kestirimler incelendiğinde her iki modelden elde edilen sonuçlar benzer olsa da standart MTK modeli sonuçlarının daha çok farklılaştığı görülmektedir. Ancak bu parametre için korelasyon ve bu kestirimlerin ortalama farklılıklarına dayalı istatistikler göz önüne alındığında, MTK ve TRMTTKT modellerinin yüksek korelasyon gösterdiği ve küçük RMSD değeriyle birbirine oldukça uyumlu olduğu belirlenmiştir. Bu bulgu Eckes (2014) ve Özdemir (2017) tarafından yapılan çalışmaların bulgularıyla paralellik göstermektedir. Ayrıca yetenek parametresine ilişkin standart hatalar MTTK modelinde daha yüksek kestirilmiştir. Literatürde de madde takımı etkisinin göz ardı edildiğinde yetenek parametresinin standart hatasının olduğundan düşük kestirildiği belirtilmektedir (Chang & Wang, 2010; Wainer, Bradlow, & Du, 2000; Wainer & Wang, 2000).

Sonuç olarak bu çalışmada standart MTK modeli ile MTTK modelinden elde edilen sonuçların birbirine oldukça yakın olduğu belirlenmiştir. Bu sonuç gerçek veri seti üzerinden yapılan çalışmalarda da bu şekildedir (Baghaei ve Ravand, 2016; Demars, 2006; Eckes, 2014; Eckes ve Baghaei, 2015; Özdemir, 2017; Yılmaz Kogar ve Kelecioğlu, 2017). Bu çalışmada bu sonucun nedeni büyük olasılıkla çalışmada kullanılan veri setinde bulunan madde takımlarının madde takımı varyanslarının düşük olmasıdır. Çünkü Glas vd. (2000) 0.50'ten düşük madde takımı etki parametrelerinin sonuçlar üzerinde göz ardı edilebilir bir etki yaptığını belirtmişlerdir. Ayrıca bu durumda 2PL veya 3PL gibi modellerin parametre tahmininin kalitesinden ödün vermeden kullanılabileceğini ifade etmişlerdir. Ancak madde takımı etkisinin yüksek olduğu belirlenen çalışmalar da bile standart MTK modelleri ve MTTK modelleri arasındaki korelasyonlar yüksek bulunmuştur (Baghaei ve Ravand, 2016; Özdemir, 2017). Ancak parametrelerden elde edilen RMSE ve standart hatalarda kısmen farklılaşmalar olduğu görülmüştür. DeMars (2006) belirttiği gibi karmaşık model daha az karmaşık modele göre biraz daha yüksek RMSE'ye yol açmıştır. Standart hatalardaki farklılıklar ise özellikle yetenek parametresinde gözlenmiştir. Bu tür farklılıklar yüksek riskli kararlar söz konusu olduğunda olumsuz sonuçlara yol açabilir (Baghaei ve Ravand, 2016). Ayrıca bu durum, test sonlandırma kriteri kişi tahminlerinin standart hatası olan bilgisayar uyarlamalı testler kullanıldığında da ciddi sorunlara yol açabilir.

Bu araştırmanın alana katkısı olmasına rağmen, daha fazla araştırma gerektiren bazı sınırlılıkları vardır. Çalışmada gerçek veriler kullanıldığı için mevcut durumun sonuçları incelenmiş ve madde takımı etki varyansının düşük olduğu kestirilmiştir. Farklı çalışmalarla bu etkilerin ne kadar yüksek olabileceği gerçek verilere dayanılarak incelenebilir. Ayrıca sadece bu etkiyi belirlemek yerine, bu etkinin yarattığı varyansın kaynağını belirlemeye yönelik çalışmalar yapılabilir. Bu amaçla, testteki her bir maddeye ulaşılabilen gerçek veriler kullanılarak madde takımının karakteristik özellikleri incelenebilir.

Mevcut çalışmada madde takımlarını ele alan modellerden biri olan 2PL MTTK modeli kullanılmıştır. MTTK modelleri, bifaktör modelinin sınırlı bir şeklidir. Bu nedenle madde takımı etkisi bifaktör modeliyle de ele alınabilir. İleride yapılacak çalışmalarda bifaktör modeli ve daha fazla parametre içeren modeller kullanılarak benzer çalışmalar yapılabilir.

Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods*

Zafer ÇEPNİ **

Hülya KELECİOĞLU ***

Abstract

In this study, differential item functioning (DIF) and differential bundle functioning (DBF) analyses of the Academic Staff and Postgraduate Education Entrance Examination Quantitative Ability Tests were carried out. Mantel-Haenszel, logistic regression, SIBTEST, Item Response Theory-Likelihood Ratio and BILOG-MG DIF Algorithm methods were used for DIF analyses. SIBTEST was the method used for DBF analyses. Data sets for the study came from an earlier application of the examination. Gender DIF analyses showed that eleven items showed DIF. Four of the items favored male applicants, where seven of them favored female applicants. In order to investigate the sources of DIF, we consulted experts. In general, the items which could be solved using routine algorithmic operations and which are presented in the algebraic, abstract format showed DIF in favor of females. The “real-life” word problems favored males. According to DBF analyses, the operations item group favored females and the word problems item group favored males.

Key Words: DIF, DBF, SIBTEST, ALES

INTRODUCTION

Large-scale tests are used to make important decisions about individuals. Large-scale exams that the Turkish community is familiar with include university entrance examinations, transition examinations for secondary education, Public Personnel Selection Examination, and Academic Personnel and Postgraduate Education Entrance Examination (Turkish acronym ALES). The first two of these exams are used for student selection. KPSS is used for staff selection and ALES is used for both student and staff selection. Over 200,000 candidates participated in ALES in 2016, which is implemented twice a year according to the information obtained from the website of the Measurement, Selection and Placement Center (Turkish acronym ÖSYM). Considering these features, ALES is one of the major large-scale exams in Turkey.

ALES consists of quantitative and verbal ability tests. Quantitative ability tests aim to measure quantitative and logical reasoning skills. The tests include items that candidates who have graduated from different bachelor's programs can answer correctly (ÖSYM, 2008). When the content of the ALES quantitative tests used in different years is examined, it is observed that the subject areas of the materials, in general, do not exceed the ninth-grade level. Content areas like trigonometry, complex numbers, limit, derivatives and integrals with which only the students in quantitative branches of high schools would be familiar are not included in the ALES quantitative tests. The difference between the quantitative 1 test and the quantitative 2 test is described as "more advanced items are used in the quantitative 2 test" (ÖSYM, 2008).

It is an indispensable requirement to present validity evidence for the large-scale examinations in which important decisions are made about candidates. One of the major threats to efficacy is item and test bias (Clauser & Mazor, 1998). For this reason, the scores should be fair to different groups taking the exams. Test fairness is not only a technical issue within the validation procedure but also an issue having

* This paper was derived from the first author's doctoral thesis titled *Detecting differential item functioning using SIBTEST, Mantel-Haenszel, logistic regression and Item Response Theory methods*.

** Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, cepni@hacettepe.edu.tr, ORCID ID: 0000-0002-8033-905X

*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:

Çepni, Z. & Kelecioğlu, H. (2021). Detecting Differential Item Functioning Using SIBTEST, MH, LR and IRT Methods. *Journal of Measurement and Evaluation in Education and Psychology*, 12(3), 267-285. doi: 10.21031/epod.988879

Received: 31.08.2021

Accepted: 24.09.2021

political, philosophical, economic, social and legal aspects (Camilli, 2006). In this framework, providing empirical evidence for test fairness is considered an important part of test development and validity studies (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014; Joint Committee on Testing Practices, 2004).

In order to understand test fairness, the concepts of item effect, statistical bias and differential item functioning (DIF) should be explained. Different performances of different groups on an item or a test is called item effect or test effect. Observation of the item or test effect does not necessarily mean that the item or test is biased (Clauser & Mazor, 1998; Millsap & Everson, 1993). If the cause for the different performances is seen as the item itself, then there is statistical bias. Here, there are differences according to groups in estimating some parameters. Statistical bias could appear in two ways. Firstly, the item parameters in the measurement model may be different for groups. This can be explained as DIF in the sense that impairment of measurement equivalency with regard to internal criteria. In the analysis of this kind of situation, an answer to the question “Does this item measure the same variable that the rest of the exam measures?” is sought. Secondly, the intercept or slope of the line used in predicting an external criterion or the standard error of the prediction may differ for different groups. This situation could be expressed as impairment of measurement equivalency with regard to external criteria or differential prediction. In an analysis of this kind of situation, the question is whether the test measures the same construct for both groups according to an external criterion (Camilli, 2006).

DIF refers to the fact that the performances of individuals from the reference and focus groups at the same level of ability are different. An item that does not exhibit DIF has the same measurement properties for reference and focus groups. In other words, for an item that does not show DIF, the likelihood of individuals with equal ability to respond to the item correctly is the same even if the individuals belong to different groups. However, if different item difficulties are observed in different groups of equal skill levels, the item exhibits DIF (Millsap & Everson, 1993). Since the DIF analyses are based on internal criteria, they assume that other validity evidence is sufficient (Clauser & Mazor, 1998). Therefore, it is generally appropriate to establish the factor structure of the tests before DIF analyses.

Although tests are often considered unidimensional, it is rare that the ability to answer an item correctly is only one. Within the multidimensionality-based DIF paradigm framework, the groups are statistically matched on the primary factor measured by the test, θ . The secondary skills required to correctly answer the item in the same paradigm are considered as η . If the groups differ on the secondary skills that the items measure, DIF is seen in these items. In other words, the reason why the item shows DIF is the difference between the groups on the secondary factor (η). There is a secondary variable (η) that is effectively functioning in a DIF item. This secondary variable, which leads to DIF, can be determined by examining the item by experts. The decision of flagging the item as biased or not is based on what the secondary variable is. If the experts see the secondary variable as an element not to be included in the construct measured by the test, the item is labelled as biased and should be removed from the test. For example, if a secondary variable such as "familiarity with hunting terms" plays a role in the analysis of any material in the test of reading skills, it may be suggested to remove the item from the test (Ackerman, 1992). If these secondary variables are deemed as integral to the construct being measured, then the item is not considered biased—only a DIF item. For example, word problems in mathematics tests may show DIF because of the effect of reading skills in their responses. Whether this DIF should be taken as bias is determined by assessing whether the reading skill is a secondary variable considered to be measured by these items. If the reading skills are a secondary variable that is desired to be measured by those items, the items are treated as DIF items only, not biased. If the undesired variables lead to DIF, the item could be considered as biased (Zumbo & Gelin, 2005). In this framework, DIF is a necessary but not sufficient condition for items to be biased (Zumbo, 1999).

As Ong, Williams and Lampranou (2011) point out, the bias decision depends on the boundaries of the target construct to be measured, and clear cut limits are not always published or easy to draw. For example, when algorithmic procedural knowledge items function in favour of female candidates, it seems that the ability to perform operations in a step-by-step and organized manner is also effective in

these items as well as general quantitative skills. Significant differences in the secondary construct between male and female candidates lead to DIF in such items. Whether or not these items will be flagged as biased would be determined by whether the ability to perform those procedures in a step-by-step and organized manner is within the target construct to be measured.

The main purpose is to eliminate item bias, which is an important threat to test validity. In this case, it is advisable to remove the item from the test if the part causing the bias in the item cannot be corrected (Ackerman, 1992, Camilli, 2006, Clauser & Mazor, 1998). Determining and eliminating item bias is used for improving test validity. In this framework, it is especially important to determine the causes of DIF as well as to detect DIF items.

Potentially biased items are detected using DIF methodology. The aim here is to identify and eliminate bias resulting from test design, content and item types among different gender, ethnicity, language, culture groups and ultimately to increase test validity (AERA, *et al.* 2014). Since potentially biased items are determined using DIF analyses, DIF detection could be considered as a step in item bias detection.

Once DIF items are identified, the variables that are the source of DIF should be examined for the decision to flag the items as biased or not (Clauser & Mazor, 1998). In DIF analyses, grouping variables can be gender, country, culture, language, socioeconomic level or ethnicity (Camilli, 2006). Important DIF sources in the cross-cultural assessments which are used for international comparisons are translation inadequacies, lack of the same reciprocal of concepts in different cultures, different levels of familiarity with different concepts from different cultures, different curricula of different countries and different teaching methods and qualifications that are emphasized by different curricula (Asil, 2010; Ercikan, 1998; Grisay, de Jong, Gebhardt, Berenzer, & Halleux-Monseur, 2007; Hambleton, Merenda, & Spielberger, 2005; Yıldırım and Berberoglu, 2009). Factors like item format, content and cognitive complexity level are among the popular gender DIF sources (Bakan Kalaycıoğlu & Berberoglu, 2010; Bakan Kalaycıoğlu & Kelecioğlu, 2011; Mendes-Barnett & Ercikan, 2006; Zumbo & Gelin, 2005).

If a DIF item is functioning in favor of a group at all levels of ability, this is called uniform DIF. The item characteristic curves determined for the two groups of such an item do not intersect. An item with intersecting characteristic curves tends to favor a group to a certain level of ability and favors the other group at higher levels of skill. This is called a non-uniform DIF (Hambleton, Swaminathan, & Rogers, 1991). Only uniform DIF items are investigated in this research because uniform DIF items favor one group more significantly and the interpretations of non-uniform DIF are more complicated (Smith & Reise, 1998).

DIF Detection Methods

In DIF determination methods, the individuals in the two groups, which are generally taken as focus and reference, are matched according to their ability estimation. For these matched groups, DIF statistics are calculated using the correct response rates of the items. A hypothesis is constructed regarding the item, saying that the item functions equivalently between the groups, and a statistical significance test is performed. However, statistical significance tests are not considered satisfactory for the interpretation of the practical significance and effect size of DIF (Camilli, 2006). Therefore, methods that provide effect size statistics may be more useful in practice. Although the methods generally give similar results to some extent, they are not in perfect agreement because they use different algorithms and different matching criteria. In addition, the cut-off points they use to flag the DIF items are different (Bakan Kalaycıoğlu & Berberoglu, 2010; Doğan & Öğretmen, 2008; Gök, Kelecioğlu & Doğan, 2010). For this reason, it is recommended that researchers and test developers use multiple methods for DIF analysis (Hambleton, 2006).

It is possible to divide DIF detection methods into two groups as (1) methods using the observed raw scores in matching of individuals and (2) methods based on Item Response Theory (IRT) (Camilli, 2006). Mantel-Haenszel (Holland & Thayer, 1988), logistics regression (Swaminathan & Rogers, 1990)

and SIBTEST (Roussos & Stout, 1996a) are among the former group. Restricted factor analysis is another method based on factor analysis that does not lend itself in either group (Oort, 1992).

Mantel-Haenszel

The Mantel-Haenszel (MH) is a DIF detection method given by Holland and Thayer (1998) in the measurement literature. In this method, the total test score is used as a matching criterion. The total test score is treated as a discrete variable in constructing the equivalent ability examinees for focus and reference groups.

For analysis, a three-dimensional matrix of size $2 \times 2 \times S$ is formed, where S is the number of ability levels being generated according to the correct and incorrect answers of the individuals from different groups. For each ability level of focus and reference groups, a data structure as shown in Table 1 is analyzed.

Table 1. Data Structure Used in Mantel-Haenszel

Group	Correct	Incorrect	Total
Reference	A_j	B_j	n_{Rj}
Focus	C_j	D_j	n_{Oj}
Total	m_{ij}	m_{oj}	T_j

A likelihood ratio is obtained by using the values in the tables for each ability level. This ratio is given in Equation 1.

$$\alpha_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \quad (1)$$

The final output of the Mantel-Haenszel algorithm is the Δ_{MH} statistic, which is -2.35 times the natural logarithm of this likelihood ratio. Since the standard error of this statistic is known, a hypothesis test can be performed using a χ^2 distribution. Negative values of the Δ_{MH} statistics indicate that the item is in favor of the reference group and the positive values indicate that the item functions in favor of the focus group. In addition, since Δ_{MH} is itself an effects size measure, it can be used to interpret the practical significance of DIF. A commonly used categorization schema has been proposed by Zieky (1993), which is shown in Table 2. Mantel-Haenszel DIF statistics can be calculated by means of EZDIF software (Waller, 1998).

Table 2. Interpretation of Mantel-Haenszel DIF Statistic

Level	Value	DIF amount
A	$ \Delta_{MH} < 1$	None or negligible
B	$1 \leq \Delta_{MH} < 1.5$	Middle
C	$ \Delta_{MH} \geq 1.5$	High

Logistics regression

Swamanithan and Rogers (1990) have shown that logistic regression (LR) can be used to detect DIF. In this method, the matching criterion is the total test score. However, unlike the Mantel-Haenszel method, it is taken as a continuous variable. Group affiliation and total test score are independent variables in the logistic regression, whereas the response to the item is a dependent variable. The mean for different groups of an item is expressed in Equation 2 in the expected value.

$$\varepsilon(Y_i | X_i, G_i) = P_i \quad (2)$$

The LR equation for uniform DIF is constructed as shown in Equation 3.

$$Z_i = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_0 + \beta_1 X_i + \beta_2 G_i \quad (3)$$

An interaction term is added to the regression equation for non-uniform DIF analysis. For a hypothesis test of whether the item being inspected exhibits uniform DIF, the fit of the above model and the fit of the model obtained by subtracting the group variable can be compared. The difference between the R^2 values of the two models, ΔR^2 , indicates an effect size used to interpret the amount of DIF (Zumbo, 1999). For the interpretation of ΔR^2 values, Zumbo and Thomas (1996) and Jodoin and Gierl (2001) proposed two separate classifications given in Table 3.

Table 3. Recommended Categories of Classification for Interpreting the ΔR^2 Values

Level	Zumbo and Thomas (1996)	Jodoin and Gierl (2001)	DIF amount
A	$\Delta R^2 < .13$	$\Delta R^2 < .035$	None or negligible
B	$.13 \leq \Delta R^2 < .26$	$.035 \leq \Delta R^2 < .070$	Middle
C	$\Delta R^2 \geq .26$	$\Delta R^2 \geq .070$	High

DIF statistics calculated by Mantel-Haenzsel and logistic regression methods are quite consistent when the index values are considered, but regarding the cut-off points used in the categoricals this consistency seems to be inadequate (Bakan Kalaycıoğlu & Berberoğlu, 2010, Doğan & Öğretmen, 2008; Gök, vd. 2010). In addition, Higaldo and Lopez-Pina (2004) tested the effectiveness of logistic regression and some other methods to detect DIF under simulation conditions and showed that only 1% of the DIF items were flagged when the cut-off point .13 is used, and 20% when .035 used. As a result of this study, it was emphasized that new criteria should be determined for the interpretation of ΔR^2 statistic. Due to this condition of the ΔR^2 statistic, Bakan Kalaycıoğlu and Kelecioğlu (2011) used the first cut-off point as .010 and the second as .020, taking into account the ΔMH DMF index. Logistic regression DIF analysis can be performed in SPSS software using the SPSS codes provided by Zumbo (1999).

SIBTEST

SIBTEST method, developed by Shealy and Stout (1993), can be used in determining statistically whether or not one item and more than one item displays DIF. The item or items for which DIF analysis is to be performed is/are included in a group and the other items are put in another group and thus, the test is divided into two parts. Matching is done with the actual scores estimated by means of the total scores on the items in the second group, and the performance of the groups which are analysed for DIF is compared (Gierl, 2005). The expected scores of the applicants in the reference (R) and focus (F) groups are identified in Equations 4 and 5- where k is the score received from DIF item or items, $P_{Rk}(t)$ and $P_{Fk}(t)$ are the ratios of t score and the applicants receiving the k scores on the items.

$$ES_R(t) = \sum_k k P_{Rk}(t) \quad (4)$$

$$ES_F(t) = \sum_k k P_{Fk}(t) \quad (5)$$

These two values are used by correcting for measuring errors in the SIBTEST. In this case, the final output of the SIBTEST method, β_u DMF index, is derived as in Equation 6.

$$\beta_u = \sum_t \left([ES_R(t) - ES_F(t)] \left[\frac{N_R(t) - N_F(t)}{N} \right] \right) \quad (6)$$

$N_R(t)$ and $N_F(t)$ values in the formula indicate the number of applicants whose matching scores are t in the reference and focus groups. Because the standard error of β_u index is known, a result of a hypothesis

test can be obtained. β_u index indicates an effect size. The classification developed by Roussos and Stout (1996b) to interpret the amount of DIF is shown in Table 4. SIBTEST can be performed by using the software called SIBTEST (Stout & Roussos, 1995).

Table 4. Classification Categories Recommended for the Interpretation of β_u Values

Groups	Values	Amount of DIF
A	$\beta_u < 0.059$	None or negligible
B	$0.059 \leq \beta_u < 0.088$	Middle
C	$\beta_u \geq 0.088$	High

SIBTEST can also test whether or not more than one item display DIF synchronically. In the same vein, β_u index also shows the amount of DIF for more than one item. Yet, no systems of classification were recommended for the evaluation of the amount of DIF when used for more than one item (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Bisanz, Bisanz, & Boughton, 2003; Ong, et al. 2011). It is possible to relatively compare the β_u statistics of both groups of items. SIBTEST is a technique based on the fact that the skills necessary for responding to an item correctly are multidimensional. In this framework, when the primary skill necessary for responding to an item correctly is taken as θ and the secondary skill as η , differentiation of the distribution of different groups on η is considered to be the source of DIF (Roussos & Stout, 1996a). SIBTEST can be used in determining the characteristics of items displaying DIF, in testing the DIF hypotheses which can be constructed beforehand and in making healthier generalisations about the sources of DIF due to the fact that SIBTEST enables one to group items and to perform DIF analysis on them (Gierl, et al. 2003; Mendes-Barnett & Ercikan, 2006).

Item Response Theory-Likelihood Ratio

As the name suggests, the item response theory likelihood ratio (IRT-LR) is an IRT-based method (Thissen, Steinberg, & Wainer, 1993). Therefore, IRT-based ability estimations, and not observed scores, are used in matching individuals. The IRT-LR analyses can be performed on IRTLRDIF software (Thissen, 2001). First, a generalised model in which item parameters are freed for both groups is constructed in DIF analysis in which IRT-LR is performed. After that, the restricted model enabling one to restrict the item parameters in the same way for both groups is constructed. $-2\log$ likelihood ratios are compared for the fit between the two models. The difference between the two models is reported as G^2 statistics.

G^2 statistics makes it possible to perform a synchronic hypothesis test about whether or not all the parameters are equal in the two groups. The G^2 value is compared with the critical value of χ^2 distribution, which is the number of parameters in the degrees of freedom IRT model, and thus the hypothesis is tested. If the synchronic hypothesis testing is found to be significant for all parameters, the G^2 value is compared with 3.84- which is the critical value of single freedom degree χ^2 distribution- for difficulty and discrimination parameters and thus, hypotheses are tested. When the G^2 value used in synchronic parameter comparisons is below 3.84, it is impossible for any parameters to be algebraically significant. For this reason, the IRTLRDIF software cannot perform the test for individual parameters in such cases (Thissen, 2001). G^2 is not an effect size statistics. It is recommended that anchor items be selected by considering the other initial IRT-LR analysis and the other DIF statistics be used in IRT-LR analyses (Wang & Yeh, 2003). Six anchor items were selected for each IRT-LR analysis in this study. The G^2 values derived from the initial application of IRT-LR method and the other DIF statistics were taken into consideration in selecting the anchor items.

BILOG-MG DMF Algorithm

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) software offers an IRT-based algorithm for DIF analyses. In this algorithm, parameters are estimated for two separate groups in a way similar to

unequal groups test matching design, and they are brought on the same scale. The difference of difficulty parameters that are brought on the same scale and the standard errors for the difference are reported (du Toit, 2003). A hypothesis test is done by dividing adjusted difficulty difference values (Δb) into standard errors. The Δb values express the effect size about the magnitude of DIF amount (Smith & Reise, 1998). Yet, there are no widely used classifications of these values. BILOG-MG algorithm allows item discrimination to differ from item to item, but it does not allow differences between groups. Therefore, it is appropriate for use only in determining and interpreting uniform DIF (Smith & Reise, 1998). It is necessary to show that IRT assumptions are satisfied prior to IRT-based DIF analyses. Therefore, unidimensionality was tested in this study prior to DIF analyses.

Although DIF analyses yield consistent results on considering the indices, it is observed that they do not determine the same items as DIF display in items on considering the cut-off points (Higaldo & Lopez-Pina, 2004). Thus, it is recommended to use more than one method in DIF analyses (Hambleton, 2006). In line with this recommendation, more than one method was used in this study to detect DIF.

Differential Bundle Functioning

Items that are probable to be biased are detected through DIF analyses. However, the causes of different functioning in different groups cannot be detected through DIF analyses. Differential bundle functioning (DBF) analyses can be used in determining the sources of DIF, and thus it becomes possible to analyse whether or not the sources of DIF are accepted into the construct intended to be measured (Ong et al., 2011). These analyses test whether or not items having certain properties function as a group. In some cases, the amount of DIF displayed by items is lower than B or C levels; but when such items come together, the effect of the item group is more remarkable and it should be taken into account (Nandakumar, 1993). DBF analysis is appropriate for analysing such situations.

DBF analyses can be performed in SIBTEST method (Roussos & Stout, 1996). In this method, item groups are formed according to their certain properties and whether or not the item groups function in different ways for different groups of students is analysed. In consequence of DBF analyses, which can be used on SIBTEST software, β_u DBF statistics are calculated for each group of items. A hypothesis test is done with the significance level of these statistics. For item groups, β_u statistics is an effect size statistics expressing the amount of DBF. But no widely used schema is available for item groups level classification (Gierl et al., 2001; Gierl, et al. 2003; Ong, et al. 2011). There are studies trying to determine the sources of DIF by doing DBF analysis on pre-determined item groups in the literature (Gierl, et al. 2001; Mendes-Barnett & Ercikan, 2006; Ryan & Chiu, 2001). Shedding light on the sources of DIF in addition to determining DIF displaying items is considered as a component of finding validity evidence (Ong, et al. 2011).

Purpose

This study aims to determine the items displaying DIF as well as item groups displaying DBF in ALES quantitative ability tests according to gender and to compare the results of differing DIF detection methods in a real data set. DIF and DBF analyses were performed for this purpose in ALES quantitative ability tests administered in Fall 2008. In this way, the target was to determine the items displaying DIF in ALES quantitative ability test and to reveal the causes for different functioning of items according to groups by using DBF analyses.

METHOD

Data Set

The raw data necessary for the study were obtained from ÖSYM. After obtaining the entire national data set from the application of ALES, the candidates who responded to at least one item correctly were

taken into consideration. The analyses were carried on the population so as to prevent the errors from being caused by sample formation. Yet, SIBTEST software can work with data sets having 7,000 participants in each group. Therefore, samples of randomly selected 13,000 applicants from quantitative test 1 and 11,000 applicants from quantitative test 2 were formed for analyses to be performed through SIBTEST by using SPSS software. The whole data set was used in data analyses apart from SIBTEST. The whole data set for quantitative test 2 and the distribution of the sample according to gender and department scores are shown in Table 5. The data set included 133,788 applicants for quantitative test 1 and 103,088 applicants for quantitative test 2. Of the applicants, 51% were female, whereas 49% were male in the quantitative test 1. The proportion was also similar in quantitative test 2. It was found that the data set chosen for SIBTEST sampling represented the data set taken as the population in terms of such variables as gender and department.

Table 5. Distributions of Scores

Gender	Quantitative 1 Test				Quantitative 2 Test			
	Whole data set		SIBTEST sample		Whole data set		SIBTEST sample	
	<i>N</i>	%	<i>n</i>	%	<i>N</i>	%	<i>n</i>	%
Female	68170	51	6629	51	53725	52	5636	51
Male	65618	49	6371	49	49363	48	5364	49
Total	133788	100	13000	100	103088	100	11000	100

Data Analysis

The data were coded by marking correct answers as 1 and marking incorrect or empty answers as 0. Prior to DIF analyses, a unidimensional measurement model was tested through confirmatory factor analysis by means of the asymptotic covariance matrix for quantitative 1 and quantitative 2 tests in order to test the unidimensionality of the data coming from the tests. PRELIS software was used in deriving asymptotic covariance matrix, whereas SIMPLIS software was used in performing the confirmatory factor analysis. Score distributions for the tests were determined and the test statistics and α coefficients were calculated. In addition to that, the item difficulties and discrimination indices for the overall test and for the sub-groups were also calculated.

Mantel-Haenszel, logistic regression, SIBTEST, IRT-LR and BILOG-MG DIF algorithm techniques were used in determining the items displaying DIF. Mantel-Haenszel analysis was done by using EZDIF software, logistic regression analysis was performed by using the codes provided by Zumbo (1999) and by using SPSS software, SIBTEST was performed by using the software carrying the same name, IRT-LR analysis was performed by using the software IRTDIF and BILOG-MG DIF analysis was performed by using the software carrying the same name. It was found that the results of almost all hypothesis tests performed with logistic regression, IRT-LR and BILOG-MG were significant. Due to the fact that the data set used was very large, the items were marked as at least middle (B level) according to at least two methods according to the classification of the effect size of Mantel-Haenszel, logistic regression and SIBTEST techniques were determined as items displaying DIF. The G^2 statistics provided by IRT-LR for the items whose indices were calculated to be very close to the cut-off scores used in the classification and the Δb statistics provided by BILOG-MG were also taken into consideration. Since there was not a schema for classifying these two techniques, the evaluation was made by comparing the other items displaying relative DIF.

Expert opinion was consulted for the causes of different functioning of DIF displaying items. Four of the eight experts included in the study held Ph.D. in measurement and evaluation while one had a doctorate degree in science education, one had a doctorate in mathematics education, one was a student of the doctorate in measurement and evaluation and one was a student of the doctorate in mathematics education. The items displaying DIF and the directions in which they displayed DIF were shown to the experts, and their opinions on the causes for the items to display DIF were obtained via open-ended questions. The forms in which the experts stated their opinions were sent through e-mails, and the experts were also interviewed face to face. Relevant literature, as well as DIF results, was taken into

consideration in DBF analyses and bundles of items having certain properties were formed. DBF analyses were conducted with the bundles of items by using SIBTEST.

Unidimensionality, test and item statistics

A unidimensional measurement model was tested with confirmatory factor analysis through asymptotic variance matrix for each of Quantitative 1 and Quantitative 2 tests so as to test the unidimensionality of the data coming from the tests. Consequently, the unidimensional measurement model was found to have an adequate model-data fit for both tests. The model-data fit statistics for the factor analysis are shown in Table 6 and the descriptive statistics for Quantitative 1 and Quantitative 2 tests are shown in Table 7. A close examination of the statistics in Table 7 demonstrates that male participants have a slightly higher average than female participants but that they have similar score heterogeneity. It is also clear that the tests slightly differ in average discrimination according to gender groups.

Item discriminant indices for Quantitative 1 and Quantitative 2 tests took on values in the 0.40-0.93 and 0.43-0.92 range. On the other hand, ALES is an examination in which a correction formula is used against accidental success, and it is thought that applicants rarely give incidental answers to the test items. Thus, an accidental success parameter was not needed in modelling the data. Therefore, a 2-parameter logistic model was chosen in the analyses of BILOG-MG DIF algorithm and in IRT-LR analyses- which were IRT-based analyses. The scatter diagrams for the item statistics of Quantitative 1 and Quantitative 2 tests are shown in Figure 1. An examination of data concerning item difficulty makes it clear that the items in the tests rank from the easiest to the most difficult in their difficulty in a wide range. It may be stated that the items in the tests generally have high discriminating power, considering item discrimination.

Table 6. Fit Indices for Confirmatory Factor Analysis

Indices	Quantitative 1	Quantitative 2
$SB\chi^2$	278754.77	342093.08
Degrees of freedom	740	740
RMSEA	0.053	0.067
SRMR	0.058	0.062
NFI	0.99	0.99
CFI	0.99	0.99

Table 7. Test Statistics

Statistics	Quant. 1	Quant. 2	Quant. 1		Quant. 2	
	Overall	Overall	Male	Female	Male	Female
Number of applicants	133788	103088	68170	65618	53725	49363
Mean	24.19	23.54	25.14	23.20	24.39	22.52
Standard deviation	11.3	11.4	11.43	11.09	11.59	11.04
Skewness	-0.48	-0.43	-0.57	-0.40	-0.54	-0.33
Kurtosis	-1.04	-1.10	-0.97	-1.07	-1.03	-1.14
Average difficulty	0.60	0.59	0.63	0.58	0.61	0.57
Average discrimination	0.75	0.76	0.77	0.73	0.78	0.73

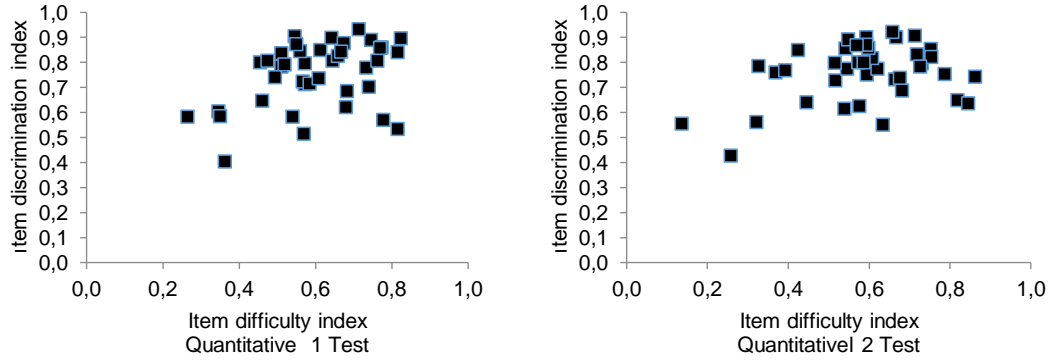


Figure 1. Scatter Diagrams for Item Statistics in Quantitative 1 and Quantitative 2 Tests

It was found that the difference between gender groups in Quantitative 1 test item difficulty was 0.18 at the maximum and the difference in item discrimination was 0.08 at the maximum. For the Quantitative 2 test, on the other hand, the difference in item difficulty was found to be 0.12 at the maximum and the difference in item discrimination was found to be 0.14 at the maximum. On estimating the item parameters within the framework of IRT separately according to applicant groups, they were not available on the same scale. Availability of item parameters in the framework of IRT on the same scale for the groups is made possible in IRT-based DIF analyses. Therefore, item statistics within the framework of IRT are given in relevant DIF analyses.

RESULTS

Findings for DIF Analyses

DIF analyses on Quantitative 1 test were performed in Mantel-Haenszel, logistic regression, SIBTEST, IRT-LR and BILOG-MG methods. The statistics considered in determining the DIF displaying items were as in the following: Δ_{MH} for MH, ΔR^2 for LR, β_u for SIBTEST, G^2 for IRT-LR and Δb for BILOG-MG DIF algorithm. Findings on DIF obtained through MH, LR and SIBTEST- which are the methods based on observed scores- are shown in Table 8. This study takes the values of 1 and 1.5 for MH, 0.010 and 0.020 for LR and 0.059 and 0.088 for SIBTEST as the criteria in determining DIF levels. The tables include only DIF displaying items. The findings for IRT-LR and BILOG DIF algorithm, which are IRT-based DIF analyses, are shown in Table 9. IRTLRDIF software uses anchor items in bringing item parameters onto the same scale. Anchor items were determined by taking other DIF statistics and item difficulty levels into consideration in this study.

Table 8. Findings for MH, LR and SIBTEST

Item no	MH		LR		SIBTEST		Advantaged group
	Δ_{MH}	Level	ΔR^2	Level	β_u	Level	
Quantitative 1							
1	-0.975		0.011	B	0.081	B	Male
6	1.383	B	0.006		-0.049		
9	-0.981		0.007		0.060	B	
10	1.998	C	0.014	B	-0.077	B	Female
11	1.211	B	0.006		-0.037		
16	1.587	C	0.011	B	-0.085	B	Female
17	1.847	C	0.015	B	-0.113	C	Female
18	1.039	B	0.003		-0.056		
20	1.436	B	0.009		-0.082	B	Female
21	-1.751	C	0.019	B	0.115	C	Male
30	-1.024	B	0.015	B	0.101	C	Male
Quantitative 2							
5	1.397	B	0.012	B	-0.061	B	Female
15	1.709	C	0.014	B	-0.085	B	Female
16	1.391	B	0.009		-0.073	B	Female
23	-1.224	B	0.006		0.065	B	Male
36	-0.915		0.004		0.059	B	

On considering the DIF statistics, which were obtained from IRT-based methods, it was found that almost all the items were marked as DIF displaying items. Therefore, the items which were marked as items having DIF according to at least two of the methods of MH, LR and SIBTEST were considered as DIF displaying items; and which groups they offered advantages was analysed. Accordingly, items 1, 21 and 30 in Quantitative 1 test and item 23 in Quantitative 2 test functioned in favour of male applicants while items 10, 16, 17 and 20 in Quantitative 1 test and items 5, 15 and 16 functioned in favour of female applicants.

Table 9. Findings for IRT-LR and BILOG-MG

Item no	MTK-OO					BILOG-MG				Advantaged group
	G^2	A_{male}	A_{female}	B_{male}	B_{female}	a	B_{male}	B_{female}	Δb	
Quantitative 1										
1	1157.1	1.16	1.11	-0.25	0.13	0.72	-0.44	-0.09	0.35	Male
10	1923.4	2.41	2.27	-0.58	-0.94	1.58	-0.72	-1.07	-0.35	Female
16	1632.3	2.09	2.02	-0.29	-0.62	1.34	-0.44	-0.77	-0.33	Female
17	2066.5	2.51	2.54	0.33	0.02	1.53	0.17	-0.16	-0.33	Female
20	1378.0	2.11	2.38	0.20	-0.07	1.31	0.04	-0.26	-0.29	Female
21	2514.3	2.21	2.41	-0.13	0.23	1.29	-0.32	0.03	0.35	Male
30	1384.9	0.80	0.68	0.65	1.40	0.46	0.42	1.02	0.60	Male
Quantitative 2										
5	845.9	1.43	1.22	-1.18	-1.77	0.83	-1.33	-1.74	-0.41	Female
15	327.4	1.93	1.84	-0.58	-0.97	1.14	-0.72	-1.07	-0.36	Female
16	807.9	2.12	1.97	-0.54	-0.85	1.24	-0.69	-0.95	-0.26	Female
23	527.5	2.93	2.78	-0.49	-0.33	1.57	-0.69	-0.47	0.21	Male

Item 1, which functioned in favour of male applicants in Quantitative 1 test, required skills related to ordering fractions. The item was presented in a way that takes too much time to solve in algorithmic methods such as equalising denominators. Therefore, applicants needed to imagine how behind the fraction is on the line of numbers according to 1 so that they could solve the problem given in the item. In this aspect, it was found that the item differed from abstract items, which could be solved in

algorithmic operations. Real-life situations were presented verbally or in the form of tables in the other three items, which displayed DIF in favour of male applicants. The applicants were required to solve the problem which was developed through real-life situations by using mathematical reasoning skills. It was apparent that the three problems involved cognitive processes more complex than algorithmic operation skills. One of those items is shown in Figure 2 below. We provide English translations of the items here. The original items in Turkish could be found in the Turkish version of this paper and in fulltext of the first author's doctoral dissertation.

In the table below, the number of people immigrating to other countries from countries A, B, C, D and E with a certain population in 2007, the number of people immigrating to these countries from foreign countries, the number of people born and died in these countries are given.

	COUNTRIES				
	A	B	C	D	E
Immigrating from	1600	4200	5000	4800	3400
Immigrating to	5400	4800	7000	1000	3800
Born	3200	5800	1300	3400	5200
Died	2000	3400	3300	3600	2600

I. The population of country C has not changed.
 II. At the end of the year, the population of country B is equal to the population of country E.
 III. There are two countries with declining population.

Given the information in the table, which of the above are definitely true?

A) Only I B) Only II C) I and II
 D) I and III E) II and III

Figure 2. An Item Displaying DIF in favour of Male Applicants

It was found that six items displaying DIF in favour of female applicants were the items which could be solved with algorithmic operations given in abstract algebraic expressions. A sample for such an item functioning in favour of female applicants is shown in Figure 3. It was also clear that another example, item 20 included in Figure 3 and which also functioned in favour of female applicants, was also a real-life problem and that it was also an item using vehicles and the concept of speed as the context. The fact that the item functioned in favour of female applicants was an unexpected situation. Almost all of the experts included in the research stated that they had expected the item to function in favour of male applicants. Yet, one of the experts said that the item was expected to function in favour of male applicants but the fact that the problem could be solved by using the equation $\text{Distance} = \text{speed} \times \text{time}$ directly and that the proportions of the distance covered by the two cars or the differences could not be used might have caused the item to function in favour of female applicants rather than male applicants. It was found in studies that real-life problems of this type functioned in favour of male applicants (Harris & Carlton, 1983; Mendes-Barnet & Ercikan, 2006). Whether or not this situation observed in ALES examinations which were in contrast to the case in the relevant literature, is a frequently observed situation that should be investigated in other DIF studies to be performed in the future. Besides, studies analysing the cognitive levels in speed-time problems and the solution strategies used by applicants of different gender groups could illuminate this point.

<p>If</p> $x + \frac{1}{x} = 3\sqrt{5},$ <p>then which of the following is equal to</p> $\left(x - \frac{1}{x}\right)^2 ?$ <p>A) 37 B) 39 C) 40 D) 41 E) 43</p>	<p>If a vehicle moving from city A to city B travels at 80 kmph, it arrives at city B 5 minutes later than it is supposed to, and if it travels at 100 kmph, it arrives 20 minutes earlier.</p> <p>How many minutes is it supposed to take for this vehicle to go to city B?</p> <p>A) 120 B) 60 C) 40 D) 30 E) 20</p>
---	--

Figure 3. The Two Items Functioning in favour of Female Applicants

Expert opinion was consulted so as to investigate the causes for DIF displaying items to function differently according to gender groups. For this purpose, DIF displaying items and the ways they displayed DIF were presented to the experts and their views on the causes for different functioning were asked. The views and the items were analysed. In accordance with the experts' views, it was concluded that all the factors likely to be the causes for DIF had remained within the mathematical/ quantitative ability construct which was intended to be measured in the tests. On considering the DIF displaying items in Quantitative1 and Quantitative 2 tests according to gender, it was found that the items which were expressed abstractly in algebraic terms and which could be solved through algorithmic operations functioned in favour of female applicants while the items which were expressed as real-life problems and which could not be solved through routine algorithmic operations were in favour of male applicants. Some of the experts included in the study stated that female applicants might have been done better than male applicants at equal ability levels due to their tendency to carry out the operations regularly and step by step. Kalaycioğlu and Kelecioğlu (2011) also reported similar findings. Accordingly, it was stated that male applicants perceived mathematics as a concept more valuable and more usable in their life than female applicants did (Fennema & Sherman, 1977). This situation might have caused male applicants to be better at practical problems taken from real life than female applicants at equal ability levels. Skills such as carrying out the operations step by step and regularly- which emerge as a factor functioning in favour of female applicants- and solving real life problems by means of mathematical models- which emerge in items functioning in favour of male applicants- can be considered as skills included in the construct which is intended to be facilitated with mathematics education and to be measured with tests.

Findings for DBF Analysis

Bundles of items likely to display DBF were formed by considering the findings coming from DIF analyses and relevant literature. Because DBF analyses were conducted after DIF analyses, initial hypotheses about the groups the bundles would function in favour of were not established; instead, only findings obtained from DBF statistics were given. The six bundles formed are described below. The bundles formed on the basis of certain properties had intersection points. That is to say, some of the items belong in more than one group.

Operations. Items expressed abstractly only in algebraic and numerical terms were grouped as operational items. It was found in this study that the majority of the items functioning in favour of female applicants were of this type. Moreover, Bakan Kalaycioğlu and Kelecioğlu (2011) also report DIF findings in favour of female applicants in some of such items. Similar findings were also reported by other researchers (Bakan Kalaycioğlu & Berberoğlu, 2010; Cohen & Ibarra, 2005; Harris & Carlton, 1983).

Word problems. This bundle of items was composed of problems that presented real-life situations, in which the data were not presented in tables or charts but were presented verbally. Such items were found among the items displaying DIF in favour of male applicants. Studies are available indicating that word

problems display DIF in favour of male applicants (Bakan Kalaycıoğlu & Berberoğlu, 2010; Harris & Carlton, 1983; Mendes-Barnett & Ercikan, 2006).

Geometry. Items requiring knowledge of geometry were included in this bundle. Contrasting research findings are available in geometry items displaying DIF according to gender (Berberoğlu, 1996; Cohen & Ibarra, 2005; Doolittle & Cleary, 1987; Mendes-Barnett & Ercikan, 2006). Geometry items did not display remarkable DIF according to gender in this study.

Analytic reasoning. It is the type of item used only in ALES among the examinations administered across Turkey. It has not been described as a subject domain in the primary or secondary school mathematics curriculum. Items of this type do not require knowledge of a special mathematical subject domain, but they can be answered by solving the given situation analytically. They can be likened to puzzles. Graduate Record Examinations (GRE), examinations similar to ALES in the USA, used to contain a sub-test of such items. Yet, GRE analytic reasoning skills test was no longer a multiple-choice test following the year 2002, and it was replaced by open-ended items measuring critical thinking and analytical writing skills (Educational Testing Service, 2007). Two examples are given for this bundle of items below in Figure 4.

Answer the following two items according to the information below.

The boxes in the above arrangement are named with the letters a, c, d, e, f, g, h. The numbers from 1 to 8 are used once and placed in the boxes, increasing both from top to bottom and from right to left. An example arrangement could be as follows.

1	a			
3	b			
5	c			
6	d			
8	7	4	2	
	e	f	g	h

As seen in this example, the numbers increase both from top to bottom and from right to left.

Given the number placed in box d is 4, what number is placed in box h?
 A) 2 B) 3 C) 5 D) 6 E) 7

Which box has the same number in all the possible arrangements?
 A) a B) b C) c D) d E) e

Figure 4. Two Examples for Items of Analytical Reasoning

Items which can be solved by trying numbers. This bundle of items contains items in which answers can be found by trying the numbers given in options or the numbers probable to be appropriate. It was seen in DIF findings according to domains that the items functioning in favour of applicants of the verbal test

had this property. Such items were also reported by Scheunemann and Grima (1997) to have functioned in favour of applicants of verbal tests. Items that could be solved by trying numbers were divided into two categories as operational items and problems.

Items of secondary education (high school) curriculum. This study found items requiring knowledge of subjects that were not included in the primary education mathematics curriculum but which were included in secondary education (high school) mathematics curriculum among the items displaying DIF in favour of applicants of the quantitative test. Such items were put under the heading of items of secondary education curriculum.

The item no and the number of items included in bundles of items and the results for DBF analyses conducted with SIBTEST are shown in Table 10. Statistical significance level was chosen as 0.01, and the groups to whose advantage the item bundles having β_u statistics functioned are shown in the same table. Operational items displayed DBF in favour of female applicants and word problems displayed DBF in favour of male applicants in both tests. This was a finding in parallel to the ones obtained in DIF analyses and in the literature (Cohen & Ibarra, 2005; Harris & Carlton, 1983; Mendes-Barnett & Ercikan, 2006). It was found that geometry items in Quantitative 1 test did not display DBF but that they functioned in favour of male applicants in Quantitative 2 test.

Table 10. Item Number and Results for DBF Analysis for the Items in the Bundles

Bundles of items	Item no		Quantitative 1		Quantitative 2	
	Quantitative 1	Quantitative 2	β_u	Advantaged group	β_u	Advantaged group
Operation	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 14, 16, 17	1, 2, 3, 4, 5, 7, 12, 13, 14, 15, 16	-0.504	Female	-0.422	Female
Word problems	15, 20, 21, 22, 24, 25, 26, 29, 30, 34, 35, 36, 37	19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 32, 33, 34, 37, 38	0.446	Male	0.611	Male
Geometry	38, 39, 40	35, 36, 39, 40	-0.032		0.126	Male
Analytic reasoning	27, 28, 31, 32, 33	29, 30, 31	0.177	Male	-0.031	
Number trying	2, 5, 9, 10, 13, 14, 24, 25, 26	1, 2, 6, 8, 11, 21, 22, 23	-0.078	Female	-0.026	
Secondary education curriculum	3, 8, 20, 39	4, 6, 8, 12, 13, 14, 19, 24, 25	-0.147	Female	-0.051	

It was apparent that this situation that did not appear in DIF analyses in which items were considered individually appeared in consequence of the combination of DIF effects at low levels in the items. Another situation in which small DIF effects combine and become remarkable was apparent in analytical reasoning items in Quantitative 1 test. Those items as a bundle also functioned in favour of male applicants. Items in which knowledge of subject areas that were not available in the primary education curriculum was effective were found to be in favour of female applicants. Female applicants with ability levels equal to male applicants in Quantitative 1 test were found to have answered more items requiring knowledge of subject areas, while male applicants were found to have been better at items of analytical reasoning which did not require knowledge of subject areas. Bundle of items that could be solved by trying numbers in Quantitative 1 test functioned in favour of female applicants, whereas DBF findings for this bundle were not found to be significant in Quantitative 2 test.

DISCUSSION and CONCLUSION

DIF analyses on the basis of gender demonstrated that 11 items had displayed DIF. Four of the items functioned in favour of male applicants, whereas seven items functioned in favour of female applicants. One of the items functioning in favour of male applicants was a problem of ordering rational numbers, while three were word problems in which real-life situations were given. Six items functioning in favour of female applicants were operational items which were given in abstract contexts and which could be solved with algorithmic operations. One item functioning in favour of female applicants, on the other hand, was a problem of speed. Findings concerning DIF according to gender were in parallel to the ones reported in previous studies (Bakan Kalaycıoğlu & Berberoğlu, 2010; Harris & Carlton, 1983; Mendes-Barnett & Ercikan, 2006). It may be generally said on the basis of DIF analysis results that the applicants of different gender groups having an equal number of correct answers in ALES Quantitative tests answered different items and that they had different answering patterns.

The results of DBF analyses demonstrated that the items displaying remarkable DIF at item level functioned in favour of groups. Operational items functioned in favour of female applicants in Quantitative 1 and Quantitative 2 tests. Word problems, on the other hand, functioned in favour of male applicants in both tests. Items that could be solved by trying numbers and the items requiring knowledge on subject areas in the secondary education curriculum functioned in favour of female applicants as a group. Analytical reasoning items, however, functioned in favour of male applicants. The fact that the final three groups of items function differently at the group level, although they did not display remarkable DIF at item level individually was the result of small DIF effects in items coming together and thus becoming more remarkable.

DIF analyses should be performed on all large-scale examinations as a routine and especially the sources of differential item functioning should be detected. Thus, efforts should be made to attain unbiasedness-an important component of the validity of tests administered. Due to the fact that different types of items can display DIF in different examinations, those analyses should be conducted for every examination in itself and thus, efforts should be made for healthy generalisations. In addition to DIF analyses according to gender, DIF analyses according to the departments of graduation should also be performed in ALES, an examination for which university graduates of differing branches apply and which is used in selecting students for post-graduate education. DIF analyses according to departments of graduation for the tests-which are the subject matter of this study- can be found in the doctoral dissertation from which this study was produced.

The fact that items display DIF does not necessarily mean that those items should not be used in tests. Yet, a group of applicants can be in a more advantageous position than others if the number of items supporting them is abundant in a test. Therefore, the number of items providing different groups with advantages could be balanced. Studies revealing the extent to which the presence of DIF displaying items in tests influences individual score differences could be performed. Besides, the effects of the availability of DIF displaying items in tests on test validity could also be investigated.

Uncovering the strategies applicants of differing groups use in solving the items and analysing the differences could be useful in detecting the source of DIF. Applicants may be asked to think aloud and to solve the items in this way. The operations applicants use on test booklets in solving the test items can also bring their strategies into the light. In addition to that, their approaches towards different types of items and their calculations can also be requested and thus, analyses can be done. Additionally, technologies monitoring applicants' eye movements and recording them while they are solving the items can also be employed for this purpose. The differences between applicants' solution strategies- how they use the tables and charts in a test item, for instance- can be analysed and thus, the sources of DIF can be detected more clearly.

This is an exploratory study concerning ALES rather than a confirmatory study testing initial hypothesis constructed beforehand. The findings obtained in this study and the DBF hypotheses to be developed by other researchers on ALES could also be tested in a confirmatory approach. The findings obtained in several studies can be generalised more effectively in this way, thus the sources of DIF can be demonstrated more clearly and they can be offered to test developers.

ΔR^2 –DIF statistics, which is used in DIF analyses along with the logistic regression method- is not adequate on its own in detecting DIF in items when cut-off points- which are commonly used in the literature- are used. Cut-off points that can be used in large-scale tests should be formed in ΔR^2 statistics by considering the first type of error and statistical power balance. Effect size classification, which can be used in IRT-based DIF analyses, should be made.

REFERENCES

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67-91.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Asil, M. (2010). *Uluslararası Öğrenci Değerlendirme Programı (PISA) 2006 öğrenci anketinin kültürler arası eşdeğerliğinin incelenmesi*. Yayınlanmamış doktora tezi, Hacettepe Üniversitesi, Ankara.
- Bakan Kalaycıoğlu, D., & Berberoğlu, G. (2010). Differential item functioning Analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment*, 20(5), 1-12.
- Bakan Kalaycıoğlu, D., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi. *Eğitim ve Bilim*, 36(161), 3-13.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport: American Council on Education & Praeger Publishers.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31-47.
- Cohen, A., & Ibarra, R. A. (2005). Examining gender-related differential item functioning using insights from psychometric and multicontext theory. In A. M. Gallagher ve J. C. Kaufman (eds.). *Gender differences in mathematics: An integrative psychological approach* (pp. 143-171). Cambridge: NY.
- Doğan, N., & Öğretmen, T. (2008). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel, ki-kare ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, 33(148), 100-112.
- Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement*, 24(2), 157-166.
- du Toit, M. (Ed.). (2003). *IRT from SSI: BILOGMG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International, Inc.
- Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, 29, 543-553.
- Educational Testing Service. (2007). *The GRE® Analytical Writing Measure: An asset in admissions decisions*. Downloaded from www.ets.org/Media/Tests/GRE/pdf/gre_aw_an_asset.pdf
- Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal*, 14(1), 51-71.
- Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice*, 24(1), 3-14.
- Gierl, M. J., Bisanz, J., Bisanz, G. L., & Boughton, K. A. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, 40(4), 281-306.
- Gierl, M., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20(2), 26–36.
- Gök, B., Kelecioğlu, H. ve Doğan, N. (2010). Değişen madde fonksiyonunu belirlemede Mantel-Haenszel ve lojistik regresyon tekniklerinin karşılaştırılması. *Eğitim ve Bilim*, 35(156), 3-16.
- Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berenzer, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3), 249-266.
- Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44(11), 182-188.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Sage Publications: California.
- Harris, A. M., & Carlton, S. T. (1983). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education*, 6(2), 137-151.

- Higaldo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.
- Holland, P. W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer ve H.I. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Erlbaum.
- Jodoin, M. G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349.
- Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Downloaded from <http://www.apa.org/science/programs/testing/fair-code.aspx>
- Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy–Stout’s test for DIF. *Journal of Educational Measurement, 30*(4), 293–312.
- Ong, Y.M., Williams, J. S., & Lamprianou, I. (2011). Exploration of the validity of gender differences in mathematics assessment using differential bundle functioning. *International Journal of Testing, 11*(3), 271-293.
- Oort, F. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*(2), 150–166.
- ÖSYM. (2008). *2008 Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) Sonbahar Dönemi Kılavuzu*. www.osym.gov.tr adresinden indirilmiştir.
- Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*(4), 355-371.
- Roussos, L. A., & Stout, W. F. (1996b). 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.
- Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*(1), 73-90.
- Scheunemann, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education, 10*(4), 299-320.
- Shealy, R., & Stout, W. F. (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*, 159–194.
- Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75*(5), 1350-1362.
- Stout, W., & Roussos, L. (1995). *SIBTEST user manual*. Urbana: University of Illinois.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.
- Thissen, D. (2001). *IRTLRDIF v.2.0.b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio tests for differential item functioning*. Downloaded from <http://www.unc.edu/~dthissen/dl.html>
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland ve H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale NJ: Erlbaum.
- Waller, N. G. (1998). EZDIF: Detection of uniform and non-uniform differential item functioning with the Mantel-Haenszel and logistic regression procedures. *Applied Psychological Measurement, 22*(4), 391.
- Wang, W., & Yeh, L. Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498.
- Yıldırım, H. H., & Berberoğlu, G. (1999). Judgemental and statistical DIF analyses of the PISA-2003 Mathematics Literacy items. *International Journal of Testing, 9*(2), 108-121.
- Zieky, M. (1993) Practical questions in the use of DIF statistics in test development. In P.W. Holland ve H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale NJ: Erlbaum.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.
- Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): *Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

- Zumbo, B. D., & Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1-23.
- Zumbo, B. D., & Thomas, D. R. (1996, October). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.

Analysis of Factors Affecting Individuals' Sources of Happiness with Multinomial Logistic Model

Kübranur ÇEBİ KARAASLAN *

Abstract

The happiness levels of individuals and their sources of happiness have been wondered a lot and researched from past to present. The aim of this study is to examine the factors that affect individuals' sources of happiness. The data set of the study was obtained from the Life Satisfaction Survey of the Turkish Statistical Institute. 9212 individuals were included in the study. In the study, chi-square independence tests were conducted to examine the relationship between the source of happiness and the independent variables included in the model, and multinomial logistic regression analysis was applied to determine the factors that may have an effect on the sources of happiness of individuals. As a result of the study, it has been determined that the factors of the individual's age, gender, marital status, educational status, satisfaction with income level, welfare level, life satisfaction, satisfaction with a social life are effective on sources of happiness. At such a time when it is clear that the coronavirus epidemic adversely affects many aspects of our lives, especially our psychology, and will leave a mark on our tomorrows, and the activities of decision-makers and policymakers are shed light through the study in order to increase the happiness of individuals and to ensure that the future will be better.

Key Words: Happiness, the economics of happiness, subjective well-being, microeconometrics, discrete choice model

INTRODUCTION

Happiness is a positive emotion that makes an individual's life meaningful and valuable (Muthuri, Senkubuge & Hongoro, 2020). Happiness, life satisfaction, subjective well-being have always been the focus of attention of researchers, especially social sciences. Long-term happiness is possible when we gain acquisitions for our values or goals (Diener, Sapyta & Suh, 1998; Pollock et al., 2015). Values and goals can have different meanings for each individual, and this situation has made it valuable to examine the factors affecting the sources of happiness of individuals and has been a source of motivation for this study. The aim of the study is to examine the factors that affect success, health, love, money, work, and other resources, which are the sources of happiness of individuals and will touch the spirit of individuals, and even societies, for decision-makers and policymakers, and the aim of this study is to be a guide that will contribute to making them happy.

In the body of literature, the concepts of subjective well-being, happiness, and life satisfaction are intertwined. In his study, Diener (2016) defined subjective well-being as a scientific term used for happiness and life satisfaction. There are many studies examining the effect of subjective well-being on different issues. As a result of the examining that Winkelmann (2005) conducted on the factors affecting the subjective well-being of individuals with the ordinal probit regression model; it has been determined that there is a "u" relationship between age and subjective well-being, unemployment negatively affects subjective well-being, and health is an important determinant of subjective well-being. Similarly, Chen and Short (2008), who investigated the effects of households on the subjective well-being of individuals, determined that subjective well-being of lonely individuals is lower, living with a close family (spouse or children) positively affects subjective well-being, health, education, and financial independence positively affects subjective well-being. Likewise, some studies examined subjective well-being with

* Asst. Prof., Erzurum Technical University Faculty of Economics and Administrative Sciences, Erzurum-Turkey, kubranur.cebi@erzurum.edu.tr, ORCID ID: 0000-0001-9288-017X

To cite this article:

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. doi: 10.21031/epod.925631

Received: 22.04.2021

Accepted: 12.09.2021

more specific titles. Carandang et al. (2020) examined the subjective well-being of individuals over the age of 60 through hierarchical regression analysis and as a result, they identified that psychological resilience is the strongest predictor of subjective well-being, and health has a positive effect on subjective well-being for both men and women. Schnepf (2010) examined gender differences in terms of subjective well-being with logistic regression analysis and found that the gender difference in subjective well-being was more dominant in post-communist countries than in OECD countries, and highly educated women had lower subjective well-being than men. Scorssolini-Comin and Santos (2011) examined the relationship between marriage and subjective well-being with multiple regression analyses and found that subjective well-being had a positive effect on marriage. Ngamaba, Panagioti, and Armitage (2017) and Bussière, Sirven, and Tessier (2021) investigated the relationship between subjective well-being and health in their studies and found that there was a positive relationship between the health status of individuals and their subjective well-being. Minarro et al. (2021), on the other hand, examined the relationship between money and subjective well-being and found that subjective well-being cannot be achieved by earning a lot of money.

Warner Wilson, who made important contributions to the field of subjective well-being in 1967, stated in his study that a happy person was "a young, healthy, well-educated, well-paid, extroverted, optimistic, worry-free, religious, married person and has high self-esteem and job satisfaction" (as cited in Diener et al., 1999). Despite the diversity in definitions of happiness, studies show that an increase in individual happiness improves not only the individual but also the community in which he or she resides (Elliot, Cullen, and Calitz, 2018). With the examination of the factors affecting people's happiness, subjective well-being, or life satisfaction, useful information can be obtained in order to reach happy individuals and, therefore happy societies. Thus, the concept of happiness should not be considered as a psychological phenomenon only and should be handled sophisticatedly. While Bülbül and Giray (2011) analyzed the relationship between sociodemographic characteristics and perception of happiness with canonical regression analysis, they determined that the happiness level of men with a job, secondary school graduates, and low incomes is in the medium and high level, Akın and Şentürk (2012) determined that although the level of happiness differed in terms of demographic characteristics, it gave basically similar results as a result of examining the factors affecting the level of happiness with ordinal logistic regression analysis. Çağlayan-Akay and Timur (2017), who investigated the factors affecting the happiness of women and men with the ordinal logistic regression model, found that economic factors were effective on happiness, and being hopeful positively affected the probability of being happy for women and men. Moyano-Diaz, Mendoza-Llanos, and Paez-Rovira (2021), on the other hand, found that loneliness and inadequate communication negatively affected people's happiness as a result of examining the socio-psychological aspects of being happy with hierarchical regression analysis.

In this study, the life satisfaction survey conducted by the Turkish Statistical Institute was used and the Discrete Choice Model, which is appropriate for the dataset, was applied and the results were presented. In the continuation of the study, first, the methodology was discussed, then the findings and the model prediction results were included. In the conclusion and evaluation part, evaluations related to the literature are presented both in terms of happiness levels and sources of happiness.

METHOD

Sample

In this study, survey data obtained through the Life Satisfaction Survey conducted by the Turkish Statistical Institute in 2019 were used. Household members aged 18 and over living within the borders of the Republic of Turkey were included in the survey. The sampling method of the research is two-stage stratified cluster sampling. In the micro data set, there are data on various subjects such as happiness, level of life satisfaction, satisfaction in basic living areas, education, health, level of hope (Turkish Statistical Institute, [TURKSTAT], 2021). A total of 9212 people who participated in the Life Satisfaction Survey in 2019 were included in this study.

Variables

The dependent variable used in the study is the sources of happiness. This variable is measured with the statement "What makes you happy the most in life? (Success; Health; Love; Job; Other)". Within the study, job, money, and other options were combined and assigned to a single category due to their low observation content. Thus, the dependent variable categories are; 1 for Success, 2 for Health, 3 for Love, 4 for Job, Money, and Other.

A literature review was conducted for the independent variables in the study. Afterward, chi-square analyzes were made, and independent variables were included in the model. In the study, sociodemographic, economic, and individual factors that may be effective on individuals' sources of happiness were taken as independent variables. Age (18-27,28-37, 38-47, 48-57, 58-67, 68 and +), gender, an education level (not finished school, primary school graduate, secondary-primary school graduate, high school graduate, college-faculty graduate, 5 or 6-year college postgraduate), marital status (married, single, widowed-divorced) variables are sociodemographic factors. Employment status of the individual (working, not working but still related to his job-not working), satisfaction with monthly income level (satisfied (very satisfied-satisfied), medium, not satisfied (not satisfied-not satisfied at all)), welfare level (low (0,1,2,3,4), medium (5), high (6,7,8,9,10)) variables are economic factors. Individual's level of happiness (happy (very happy-happy), moderate, not happy (unhappy-very unhappy)), those who make happy (self, children-spouse, whole family-niece-granddaughter, other-friends) life satisfaction (not satisfied (0,1,2,3,4), moderate (5), satisfied (6,7,8,9,10)), satisfaction with health (satisfied (very satisfied-satisfied), moderate, dissatisfied (not satisfied) not satisfied at all)), satisfaction with the education he received (satisfied (very satisfied-satisfied) medium, not satisfied (not satisfied-not satisfied at all), not educated)), satisfaction with his social life (satisfied (very satisfied-satisfied), moderate, dissatisfied (not satisfied at all)), hope (very hopeful-hopeful, hopeless-very hopeless), past comparison (improved, same, regressed, no idea), future comparison (will improve, same, regressed, no idea) variables are individual factors.

Data Analysis

Microsoft Excel was used to make the data suitable for analysis, SPSS 20 for chi-square independence tests and Stata 14.1 for multinomial logistic regression analysis were used.

The discrete choice models, which are the backbone of empirical analysis for many fields, including economics, psychology, transportation, public policy, are used to estimate the probability of choosing an alternative under the assumption that decision-makers will maximize utility among finite alternatives (Ben-Akiva & Bierlaire, 1999; Garrow, 2016; Newman, Lurkin & Garrow, 2018). The multinomial logistic regression model, which is one of the discrete choice models, is applied when the dependent variable contains three or more categories without being subjected to an order (Koppelman & Wen, 1998).

Since the dependent variable of the study is the sources of happiness of individuals, multinomial logistic regression model, which is one of the discrete choice models, was used in the analysis of the data due to the categorical nature of the dependent variable

In the study, firstly, the frequencies and percentages of the individuals participating in the study were calculated according to their sources of happiness. Afterward, chi-square independence tests were conducted to examine the relationship between the source of happiness and the independent variables included in the model, and multinomial logistic regression analysis was applied to determine the factors that may have an effect on the sources of happiness of individuals.

RESULTS

Descriptive Statistics and Chi-square Tests

The independent variables that may be effective on the happiness sources of individuals within the study and the frequency values of their categories are shown in Table 1.

Table 1. Frequencies and Percentages of Sociodemographic, Economic and Individual Factors According to Individuals' Sources of Happiness

Variables	f(%)	Sources of Happiness			
		Success	Health	Love	Job, Money, and Other
Sociodemographic Indicators					
<i>Age</i>					
18-27	1589(17.2)	319(41.9)	881(13.5)	252(18.5)	137(24)
28-37	1844(20)	155(20.3)	1238(19)	327(24)	124(21.7)
38-47	1979(21.5)	147(19.3)	1421(21.8)	304(22.4)	107(18.7)
48-57	1586(17.2)	78(10.2)	1202(18.4)	225(16.5)	81(14.2)
58-67	1183(12.8)	43(5.6)	917(14.1)	140(10.3)	83(14.5)
68 and more	1031(11.2)	20(2.6)	859(13.2)	112(8.2)	40(7)
<i>Gender</i>					
Male	4226(45.9)	455(59.7)	2845(43.6)	590(43.4)	336(58.7)
Female	4986(54.1)	307(40.3)	3673(56.4)	770(56.6)	236(41.3)
<i>Marital Status</i>					
Never Married	1597(17.3)	371(48.7)	842(12.9)	220(16.2)	164(28.7)
Married	6702(72.8)	358(47)	4967(76.2)	1023(75.2)	354(61.9)
Divorced-Widowed	913(9.9)	33(4.3)	709(10.9)	117(8.6)	54(9.4)
<i>Educational Status</i>					
Not Finish A School	1260(13.7)	19(2.5)	1019(15.6)	142(10.4)	80(14)
Primary School	2982(32.4)	132(17.3)	2266(34.8)	412(30.3)	172(30.1)
Secondary School	1385(15)	115(15.1)	955(14.7)	221(16.3)	94(16.4)
High School	1827(19.8)	262(34.4)	1166(17.9)	265(19.5)	134(23.4)
College. License	1580(17.2)	210(27.6)	1007(15.4)	282(20.7)	81(14.2)
Postgraduate for 5 or 6-Year Faculty	178(1.9)	24(3.1)	105(1.6)	38(2.8)	11(1.9)
Economic Indicators					
<i>Employment Status</i>					
Working	3890(42.2)	395(51.8)	2615(40.1)	619(45.5)	261(45.6)
Not Working	5322(57.8)	367(48.2)	3903(59.9)	741(54.5)	311(54.4)
<i>Satisfaction with Income Level</i>					
Satisfied	3755(40.8)	329(43.2)	2680(41.1)	564(41.5)	182(31.8)
Moderate	2102(22.8)	186(24.4)	1531(23.5)	313(23)	72(12.6)
Not Satisfied	3355(36.4)	247(32.4)	2307(35.4)	483(35.5)	318(55.6)
<i>Welfare Level</i>					
Low	3782(41.1)	320(42)	2661(40.8)	506(37.2)	295(51.6)
Moderate	2492(27.1)	171(22.4)	1801(27.6)	393(28.9)	127(22.2)
High	2938(31.9)	271(35.6)	2056(31.5)	461(33.9)	150(26.2)
Individual Indicators					
<i>Happiness Level</i>					
Happy	4952(53.8)	334(43.8)	3661(56.2)	759(55.8)	198(34.6)
Moderate	3103(33.7)	322(42.3)	2129(32.7)	456(33.5)	196(34.3)
Not happy	1157(12.6)	106(13.9)	728(11.2)	145(10.7)	178(31.1)
<i>Those Who Make Happy</i>					

Self	313(3.4)	78(10.2)	160(2.5)	39(2.9)	36(6.3)
Children and Spouse	1658(18)	111(14.6)	1155(17.7)	273(20.1)	119(20.8)
Mother and Father	214(2.3)	55(7.2)	103(1.6)	25(1.8)	31(5.4)
Whole Family	6914(75.1)	492(64.6)	5048(77.4)	1009(74.2)	365(63.8)
Other	113(1.2)	26(3.4)	52(0.8)	14(1)	21(3.7)
<i>Life Satisfaction</i>					
Satisfied	2696(29.3)	215(28.2)	1890(29)	327(24)	264(46.2)
Moderate	2156(23.4)	161(21.1)	1574(24.1)	290(21.3)	131(22.9)
Not Satisfied	4360(47.3)	386(50.7)	3054(46.9)	743(54.6)	177(30.9)
<i>Satisfaction with Health</i>					
Satisfied	6173(67)	570(74.8)	4270(65.5)	966(71)	367(64.2)
Moderate	1817(19.7)	125(16.4)	1341(20.6)	249(18.3)	102(17.8)
Not Satisfied	1222(13.3)	67(8.8)	907(13.9)	145(10.7)	103(18)
<i>Satisfaction with the Education Received</i>					
Satisfied	5057(54.9)	443(58.1)	3547(54.4)	780(57.4)	287(50.2)
Moderate	1337(14.5)	122(16)	925(14.2)	202(14.9)	88(15.4)
Not Satisfied	2239(24.3)	192(25.2)	1581(24.3)	313(23)	153(26.7)
Did not Receive Education	579(6.3)	5(0.7)	465(7.1)	65(4.8)	44(7.7)
<i>Satisfaction with Social Life</i>					
Satisfied	4419(48)	389(51)	3128(48)	681(50.1)	221(38.6)
Moderate	2013(21.9)	141(18.5)	1497(23)	270(19.9)	105(18.4)
Not Satisfied	2780(30.2)	232(30.4)	1893(29)	409(30.1)	246(43)
<i>Hope</i>					
Hopeful	6483(70.4)	508(66.7)	4657(71.4)	993(73)	325(56.8)
Hopeless	2729(29.6)	254(33.3)	1861(28.6)	367(27)	247(43.2)
<i>Past Comparison</i>					
Improved	2644(28.7)	276(36.2)	1811(27.8)	426(31.3)	131(22.9)
Same	2615(28.4)	177(23.2)	1944(29.8)	355(26.1)	139(24.3)
Regressed	3822(41.5)	304(39.9)	2651(40.7)	570(41.9)	297(51.9)
No idea	131(1.4)	5(0.7)	112(1.7)	9(0.7)	5(0.9)
<i>Future Comparison</i>					
Will Improve	2603(28.3)	292(38.3)	1731(26.6)	434(31.9)	146(25.5)
Same	2911(31.6)	180(23.6)	2186(33.5)	399(29.3)	146(25.5)
Will Regress	2835(30.8)	236(31)	1967(30.2)	407(29.9)	225(39.3)
No idea	863(9.4)	54(7.1)	634(9.7)	120(8.8)	55(9.6)

According to the findings, 21.5% of individuals are in the 38-47 age range and 54.1% are women. Most of the individuals included in the study (72.8%) are married. While 13.7% of individuals have not completed school, 19.1% are university graduates and 57.8% are not working. While 40.8% of the individuals are satisfied and very satisfied with the monthly income of the household, the welfare level of 41.1% is below the average. It has been determined that 53.8% of individuals are happy and very happy, 75.1% are made happy by all family members, 47.3% are satisfied with their lives, 67% are satisfied and very satisfied with their health, 54.9% of them are satisfied and very satisfied with the education they have received, 48% are satisfied and very satisfied with their social life, 70.4% are hopeful for their future, 41.5% have a deteriorated financial and moral situation compared to 5 years ago, 31.6% of them stated that their situation would generally remain the same for the next 5-year period.

Table 2. Chi-square Independence Tests of Sociodemographic, Economic and Individual Factors According to Individuals' Sources of Happiness

Variables	χ^2	<i>Degree of Freedom</i>
<i>Sociodemographic Indicators</i>		
<i>Age</i>		
18-27	526.09 ^a	15
28-37		
38-47		
48-57		
58-67		
68 and more		
<i>Gender</i>		
Male	113.305 ^a	3
Female		
<i>Marital Status</i>		
Never Married	672.09 ^a	6
Married		
Divorced-Widowed		
<i>Educational Status</i>		
Not Finish A School	353.109 ^a	15
Primary School		
Secondary School		
High School		
College, License		
Postgraduate for 5 or 6-Year		
Faculty		
<i>Economic Indicators</i>		
<i>Employment Status</i>		
Working	49.452 ^a	3
Not Working		
<i>Satisfaction with Income Level</i>		
Satisfied	104.363 ^a	6
Moderate		
Not Satisfied		
<i>Welfare Level</i>		
Low	44.998 ^a	6
Moderate		
High		
<i>Individual Indicators</i>		
<i>Happiness Level</i>		
Happy	251.683 ^a	6
Moderate		
Not Happy		
<i>Those Who Make Happy</i>		
Self	360.907 ^a	12
Children and Spouse		
Mother and Father		
Whole Family		
Other		
<i>Life Satisfaction</i>		
Dissatisfied	124.532 ^a	6
Moderate		
Satisfied		
<i>Satisfaction with Health</i>		

	Satisfied	52.339 ^a	6
	Moderate		
	Dissatisfied		
<i>Satisfaction with the Education Received</i>			
	Satisfaction	62.784 ^a	9
	Moderate		
	Dissatisfied		
	Did not Receive Education		
<i>Satisfaction with Social Life</i>			
	Satisfied	60.456 ^a	6
	Moderate		
	Dissatisfied		
<i>Hope</i>			
	Hopeful	63.597 ^a	3
	Hopeless		
<i>Past Comparison</i>			
	Improved	75.121 ^a	9
	Same		
	Regressed		
	No idea		
<i>Future Comparison</i>			
	Will Improve	94.193 ^a	9
	Same		
	Will Regress		
	No idea		

^a $p < .01$

According to the probe values of the chi-square independence tests in Table 2, it has been determined that there are statistically significant relationships between individuals' sources of happiness and sociodemographic, economic, and individual indicators.

Model Estimation

In the study, a multinomial logistic regression model was used to determine the factors that affect individuals' sources of happiness. An important assumption of multinomial logistic regression analysis is the assumption of independence of irrelevant alternatives (Vijverberg, 2011). The assumption of independence of irrelevant alternatives means that the relative probabilities of each pair of alternatives are independent of the presence or absence of all other alternatives. Violation of this assumption leads to incorrect estimates (Greene, 2002; Koppelman and Wen, 1998). Small-Hsiao test was used to test this assumption. The results of the independence test of irrelevant alternatives of the multinomial logistic regression model are given in Table 3.

Table 3. Small-Hsiao Test Results

Dependent Variable	lnL(full)	lnL(omit)	X^2	Degree of Freedom	$P > X^2$
Success	-2714.702	-2682.510	64.384	82	0.924
Health	-1244.875	-1204.693	80.364	82	0.530
Love	-2012.036	-1972.191	79.690	82	0.552
Job, Money or other	-2938.609	-2901.638	73.942	82	0.725

H_0 : Rates are independent of other alternatives.

H_1 : Rates are not independent of other alternatives.

With reference to Table 2, it is concluded that the H_0 hypothesis cannot be rejected for categories such as success, health, love, work, money, and other categories that are sources of happiness. Thus, the assumption of independence of irrelevant alternatives is provided. Another assumption of the multinomial logistic regression model is that there is no multicollinearity between the independent variables. Because of this, variance inflation factors (vif) were examined. The variance inflation factor being less than 5 indicates that there is no multicollinearity (Alkan & Abar, 2020). All of the variance inflation factors are less than 5 and there are no independent variables with multicollinearity problems in the study.

The estimation results of the multinomial logistic regression model are given in Table 4. In the model, the "health" category of the dependent variable was taken as the reference category.

Table 4. Multinomial Logistic Model Estimation Results

Variables	Success		Love		Job, Money, and Other		Vif
	β	Std. Error	β	Std. Error	β	Std. Error	
Sociodemographic Indicators							
<i>Age (reference: 18-27)</i>							
28-37	-0.498 ^a	0.129	-0.152	0.109	-0.151	0.157	2.29
38-47	-0.390 ^a	0.146	-0.333 ^a	0.118	-0.382 ^b	0.178	2.77
48-57	-0.718 ^a	0.172	-0.405 ^a	0.126	-0.527 ^a	0.193	2.65
58-67	-0.911 ^a	0.207	-0.541 ^a	0.142	-0.185	0.200	2.45
68 and more	-1.355 ^a	0.275	-0.605 ^a	0.159	-0.825 ^a	0.246	2.66
<i>Gender (reference: male)</i>							
Female	-0.540 ^a	0.093	0.065	0.072	-0.610 ^a	0.106	1.41
<i>Marital Status (reference: married)</i>							
Never Married	0.981 ^a	0.123	-0.068	0.108	0.668 ^a	0.149	1.90
Divorced-Widowed	-0.059	0.199	-0.014	0.117	-0.001	0.169	1.25
<i>Educational Status (reference: not finish a school)</i>							
Primary School	0.602 ^b	0.296	0.143	0.141	-0.076	0.206	4.15
Secondary School	0.628 ^b	0.306	0.168	0.156	-0.256	0.230	3.16
High School	1.152 ^a	0.299	0.148	0.155	-0.106	0.224	3.82
College, License	1.191 ^a	0.302	0.354 ^b	0.157	-0.378	0.238	3.62
Postgraduate for 5 or 6-Year Faculty	1.416 ^a	0.380	0.666 ^a	0.241	-0.014	0.391	1.39
Economic Indicators							
<i>Employment Status (reference: not working)</i>							
Working	-0.098	0.094	0.053	0.074	0.012	0.106	1.48
<i>Satisfaction with Income Level (reference: moderate)</i>							
Satisfied	-0.028	0.111	-0.026	0.083	0.457 ^a	0.153	1.89
Not Satisfied	-0.222 ^c	0.118	0.039	0.086	0.725 ^a	0.146	1.92
<i>Welfare Level (reference: low)</i>							
Moderate	-0.165	0.112	0.098	0.080	-0.010	0.123	1.42
High	-0.078	0.113	-0.023	0.086	0.238 ^c	0.134	1.80
Individual Indicators							
<i>Happiness Level (reference: moderate)</i>							
Happy	-0.374 ^a	0.097	-0.103	0.072	-0.293 ^b	0.116	1.46
Not Happy	-0.074	0.142	0.057	0.115	0.625 ^a	0.129	1.44
<i>Those Who Make Happy (reference: whole family)</i>							
Self	1.024 ^a	0.164	0.125	0.188	0.783 ^a	0.207	1.08
Children and Spouse	0.489 ^a	0.122	0.199 ^b	0.080	0.462 ^a	0.120	1.10
Mother and Father	0.677 ^a	0.188	0.009	0.234	0.708 ^a	0.231	1.09
Other	0.854 ^a	0.267	0.213	0.308	1.257 ^a	0.282	1.03
<i>Life Satisfaction (reference: moderate)</i>							

Not Satisfied	-0.037	0.125	-0.075	0.095	0.170	0.127	1.86
Satisfied	0.023	0.116	0.249 ^a	0.084	-0.234 ^c	0.135	1.96
<i>Satisfaction with Health (reference: moderate)</i>							
Satisfied	-0.027	0.116	0.044	0.083	0.104	0.127	1.66
Not Satisfied	-0.053	0.170	-0.082	0.117	0.098	0.157	1.55
<i>Satisfaction with the Education Received (reference: moderate)</i>							
Satisfied	-0.013	0.120	-0.061	0.091	-0.102	0.136	2.32
Not Satisfied	0.154	0.135	-0.070	0.103	-0.156	0.150	2.15
Did not Receive Education	-0.761	0.539	-0.037	0.198	0.204	0.271	2.16
<i>Satisfaction with Social Life (reference: moderate)</i>							
Satisfied	0.196 ^c	0.116	0.146 ^c	0.084	0.033	0.133	1.88
Not Satisfied	0.351 ^a	0.125	0.186 ^b	0.091	0.182	0.133	1.88
<i>Hope (reference: hopeless)</i>							
Hopeful	-0.175 ^c	0.103	0.078	0.079	-0.089	0.111	1.39
<i>Past Comparison (reference: same)</i>							
Improved	0.149	0.122	0.061	0.092	-0.051	0.147	1.93
Regressed	0.060	0.122	0.167 ^c	0.088	0.056	0.131	2.09
No idea	-0.257	0.512	-0.638 ^c	0.368	-0.770	0.503	1.22
<i>Future Comparison (reference: same)</i>							
Will Improve	0.394 ^a	0.120	0.153 ^c	0.089	0.261 ^c	0.142	1.86
Will Regress	0.183	0.128	0.025	0.093	0.146	0.136	2.02
No Idea	0.378 ^b	0.179	0.148	0.121	0.387 ^b	0.180	1.39
Cons.	-2.679	0.381	-1.905	0.238	-2.710	0.354	
Log-likelihood	-7693.7222		<i>P</i>				0.000
AIC	15633.444		<i>N</i>				9212
BIC	16510.221						

^a*p*<.01; ^b*p*<.05; ^c*p*<.10

The estimated multinomial logistic regression model was found to be statistically significant ($P < 0.000$). According to the results of the multinomial logistic model given in Table 4, success for the source of happiness; individual's age (28-37, 38-47, 48-57, 58-67, 68, and more), gender, marital status (never married), educational status (primary, secondary, high school, college-bachelor, postgraduate-5 or 6 year faculty), satisfaction with income level (not satisfied), level of happiness (happy), those who make the individual happy (self, children and spouse, mother and father, other), social life satisfaction (satisfied, not satisfied), hope, future comparison (will develop, no idea) variables were found to be statistically significant.

Love for the source of happiness; individual's age (38-47, 48-57, 58-67, 68 and more), educational status (college-bachelor, postgraduate-5 or 6 year faculty), those who make the individual happy (children and spouse), life satisfaction (satisfied), social life satisfaction (satisfied, not satisfied), past comparison (regressed, no idea) future comparison (will improve) variables were found to be statistically significant.

For job money and other sources of happiness; individual's age (38-47, 48-57, 68 and more), gender, marital status (never married), satisfaction with income level (satisfied, dissatisfied), welfare level (high), happiness level (happy, not happy), happy (self, children and spouse, mother and father, other), life satisfaction (satisfied), future comparison (no idea) variables were found to be statistically significant.

As a result of the model estimation, the independent variables will be interpreted with the help of marginal effects. Table 5 shows the marginal effects and standard errors of factors affecting individuals' sources of happiness.

Table 5. Multinomial Logistic Model Marginal Effects

Variables	Success		Health		Love		Job, Money, and Other	
	ME	Std. Error	ME	Std. Error	ME	Std. Error	ME	Std. Error
Sociodemographic Indicators								
<i>Age (reference: 18-27)</i>								
28-37	-0.415 ^a	0.114	0.083 ^a	0.029	-0.069	0.087	-0.068	0.143
38-47	-0.272 ^b	0.129	0.117 ^a	0.031	-0.215 ^b	0.096	-0.265	0.163
48-57	-0.557 ^a	0.155	0.161 ^a	0.032	-0.244 ^b	0.104	-0.366 ^b	0.178
58-67	-0.742 ^a	0.190	0.169 ^a	0.034	-0.372 ^a	0.119	-0.016	0.182
68 and more	-1.121 ^a	0.259	0.234 ^a	0.034	-0.370 ^a	0.134	-0.590 ^b	0.230
<i>Gender (reference: male)</i>								
Female	-0.467 ^a	0.084	0.073 ^a	0.017	0.138 ^b	0.060	-0.537 ^a	0.098
<i>Marital Status (reference: married)</i>								
Never Married	0.847 ^a	0.107	-0.133 ^a	0.027	-0.201 ^b	0.092	0.535 ^a	0.134
Divorced-Widowed	-0.053	0.186	0.006	0.026	-0.009	0.098	0.005	0.158
<i>Educational Status (reference: not finish a school)</i>								
Primary School	0.558 ^b	0.283	-0.044	0.030	0.099	0.121	-0.120	0.189
Secondary School	0.590 ^b	0.291	-0.038	0.033	0.130	0.133	-0.294	0.211
High School	1.065 ^a	0.284	-0.088 ^a	0.033	0.060	0.133	-0.194	0.205
College, License	1.082 ^a	0.287	-0.108 ^a	0.034	0.246 ^c	0.133	-0.487 ^b	0.220
Postgraduate for 5 or 6-Year Faculty	1.205 ^a	0.348	-0.211 ^a	0.068	0.455 ^b	0.193	-0.225	0.360
Economic Indicators								
<i>Employment Status (reference: not working)</i>								
Working	-0.098	0.085	0.000	0.017	0.052	0.062	0.011	0.098
<i>Satisfaction with Income Level (reference: moderate)</i>								
Satisfied	-0.044	0.100	-0.015	0.019	-0.041	0.070	0.441 ^a	0.145
Not Satisfied	-0.249 ^b	0.106	-0.028	0.020	0.011	0.072	0.697 ^a	0.137
<i>Welfare Level (reference: low)</i>								
Moderate	-0.166	0.102	-0.001	0.019	0.097	0.067	-0.011	0.115
High	-0.083	0.102	-0.006	0.020	-0.029	0.072	0.232 ^c	0.123
Individual Indicators								
<i>Happiness Level (reference: moderate)</i>								
Happy	-0.311 ^a	0.088	0.063 ^a	0.017	-0.040	0.061	-0.231 ^b	0.109
Not Happy	-0.125	0.127	-0.052 ^c	0.028	0.005	0.095	0.573 ^a	0.115
<i>Those Who Make Happy (reference: whole family)</i>								
Self	0.839 ^a	0.134	-0.185 ^a	0.048	-0.060	0.156	0.598 ^a	0.182
Children and Spouse	0.387 ^a	0.109	-0.102 ^a	0.021	0.097	0.066	0.360 ^a	0.111
Mother and Father	0.561 ^a	0.158	-0.116 ^b	0.052	-0.107	0.196	0.592 ^a	0.201
Other	0.624 ^a	0.220	-0.229 ^a	0.084	-0.016	0.252	1.028 ^a	0.230
<i>Life Satisfaction (reference: moderate)</i>								
Not Satisfied	-0.036	0.113	0.001	0.020	-0.074	0.081	0.171	0.116
Satisfied	-0.003	0.104	-0.026	0.020	0.222 ^a	0.071	-0.260 ^b	0.126
<i>Satisfaction with Health (reference: moderate)</i>								
Satisfied	-0.037	0.105	-0.011	0.019	0.034	0.070	0.093	0.119
Not Satisfied	-0.043	0.154	0.010	0.026	-0.072	0.100	0.108	0.145
<i>Satisfaction with the Education Received (reference: moderate)</i>								
Satisfied	0.004	0.108	0.017	0.021	-0.044	0.076	-0.086	0.125
Not Satisfied	0.161	0.122	0.007	0.024	-0.063	0.086	-0.149	0.138
Did not Receive Education	-0.726	0.515	0.035	0.047	-0.002	0.167	0.239	0.247

<i>Satisfaction with Social Life (reference: moderate)</i>								
Satisfied	0.158	0.106	-0.038 ^b	0.018	0.108	0.071	-0.004	0.124
Not Satisfied	0.285 ^b	0.114	-0.067 ^a	0.021	0.120	0.077	0.115	0.123
<i>Hope (reference: hopeless)</i>								
Hopeful	-0.165 ^c	0.093	0.009	0.018	0.087	0.067	-0.080	0.103
<i>Past Comparison (reference: same)</i>								
Improved	0.131	0.110	-0.018	0.021	0.043	0.077	-0.069	0.136
Regressed	0.027	0.111	-0.033	0.020	0.134 ^c	0.074	0.023	0.121
No idea	-0.131	0.473	0.126 ^b	0.055	-0.513	0.336	-0.644	0.482
<i>Future Comparison (reference: same)</i>								
Will Improve	0.323 ^a	0.108	-0.071 ^a	0.021	0.083	0.075	0.190	0.132
Will Regress	0.157	0.116	-0.026	0.020	-0.001	0.078	0.120	0.127
No Idea	0.301 ^c	0.161	-0.077 ^b	0.030	0.071	0.102	0.310 ^c	0.166

^a $p < .01$; ^b $p < .05$; ^c $p < .10$

According to the multinomial logistic regression model given in Table 5, for the source of success and happiness: being 68 years old or older reduces the probability of being happy with success by 112.1% compared to the reference group. Female individuals are 46.7% less likely to be happy with success than male individuals. Individuals who have never been married are 84.7% more likely to be happy with success than married individuals. The fact that individuals are postgraduates of 5 or 6 years of faculty increases the probability of being happy with success by 120.5% compared to the reference group. Individuals who are not satisfied with their income level are 24.9% less likely to be happy with success than the reference group. Individuals who are happy with their lives as a whole are 31.1% less likely to be happy with success than the reference group. Individuals who are made happy in their lives by their mothers and fathers are 56.1% more likely to be happy with success than the reference group. Individuals who are not satisfied with their social life are 28.5% more likely to be happy with success than the reference group. Individuals who are hopeful about their own future are 16.5% less likely to be happy with success than the reference group. Individuals who think that their situation will improve in the next 5 years are 32.3% more likely to be happy with success than the reference group.

Health for the source of happiness: Individuals aged 68 and above increase the probability of being happy with health by 23.4% compared to the reference group. Female individuals are 7.3% more likely to be happy with health than male individuals. Individuals who have never been married are 13.3% less likely to be happy with their health than married individuals. Being a postgraduate-5 or 6 year faculty for individuals reduces the probability of being happy with health by 21.1% compared to the reference group. Individuals who are made happy in their lives by their mothers and fathers are 11.6% less likely to be happy with health than the reference group. Individuals who are not satisfied with their social life are 6.7% less likely to be happy with their health than the reference group. Individuals who think that their general condition will improve in the next 5 years are 7.1% less likely to be happy with their health than the reference group.

Love for the source of happiness: Individuals aged 68 and over decrease the probability of being happy with love by 37% compared to the reference group. Female individuals are 13.8% more likely to be happy with love than male individuals. Individuals who have never been married are 20.1% less likely to be happy with love than married individuals. The fact that individuals are postgraduates of 5 or 6 years of faculty increases the probability of being happy with love by 45.5% compared to the reference group. Individuals who are satisfied with their lives are 22.2% more likely to be happy with love than the reference group.

For job, money, and other sources of happiness: Individuals aged 68 and above reduce the probability of being happy with a job, money, and other sources of happiness by 59% compared to the reference group. Female individuals are 53.7% less likely to be happy with a job, money, and other sources of happiness than male individuals. Individuals who have never been married are 53.5% more likely to be happy with a job, money, and other sources of happiness than married individuals. Being a postgraduate

of college-bachelor for the individuals decreases the probability of being happy with job, money, and other sources of happiness by 48.7% compared to the reference group. Individuals who are satisfied with their income level are 44.1% more likely to be happy with a job, money, and other sources of happiness than the reference group. Individuals with a high level of well-being are 23.2% more likely to be happy with a job, money, and other sources of happiness than the reference group. Individuals who are happy with their lives as a whole are 23.1% less likely to be happy with a job, money, and other sources of happiness than the reference group. Individuals who are made happy in their lives by their mothers and fathers are 59.2% more likely to be happy with a job, money, and other sources of happiness than the reference group. Individuals who are satisfied with their lives are 26% less likely to be happy with a job, money, and other sources of happiness than the reference group.

DISCUSSION and CONCLUSION

The happiness of individuals brings together happy societies and as a natural result, a peaceful environment occurs. In such a system, it may be possible to achieve more effective outputs with less effort for decision-makers on many vital issues from the economy to health and from education to defense. For this reason, happiness should be considered multidimensional and perhaps more emphasis should be placed on interdisciplinary studies in this regard. The happiness of individuals is affected by many factors, especially demographic and economic factors. In this study, demographic, economic, and individual factors that are effective on individuals' sources of happiness were first investigated with chi-square independence tests and then multinomial logistic regression model, which is the discrete choice model.

As a result of the study, while the happiest individuals with success are young, those who are least happy are over 68 years of age. It is possible to say that the probability of being happy because of success decreases as age increases. Parallel to this result, while the probability of being happy with money and other sources of happiness is higher in young people, it decreases after the middle-ages. In the literature, Selim (2008) determined in his study that compared to individuals in all age groups, individuals in the 18-30 age group believe more that power, job, success, money, and love bring happiness. Success is a more important source of happiness for young individuals who have a dynamic career plan compared to older individuals who have completed their career plans. In addition, this may be related to the fact that younger individuals are less satisfied with their lives compared to older individuals. Likewise, Fernández-Ballesteros, Zamarrón, and Ruiz (2001) and Peterson, Park, and Seligman (2005) determined in their studies that young individuals are less satisfied with their lives compared to the elderly. In addition to this, there are also studies in the literature that found that age affects happiness negatively (Atay, 2012; Chen & Short, 2008; Ekici & Koydemir, 2013). Young people are the most likely to be happy with love, and this probability decreases as age increases. This may be related to the fact that young individuals experience emotions such as love more intensely.

Individuals most likely to be happy with health are 68 years and older, and as the age increases, the probability of being happy with health increases. As age increases, the probability of facing health problems is higher. Thus, older individuals care more about health compared to young individuals, and they know the value of health more. Likewise, Bussière et al. (2021) found that the value given to health differs with age, and that aging increases the effect of health on subjective well-being for individuals and strengthens the relationship between them. In addition to this, when it is looked at from another point of view, health has a very important share in the happiness of individuals whether old or young without making discrimination. There are studies supporting this argument in the literature (Akın & Şentürk, 2012; Bussière et al., 2021; Carandang et al., 2020; Fernández-Ballesteros et al., 2001; Çebi-Karaaslan, Çalmaşur, & Emre-Aysin, 2021; Larson, 1978; Selim, 2008).

Compared to men, women are less likely to be happy with success, job, money, and other sources of happiness, but more likely to be happy with health and love. This may be related to the fact that women are more emotional than men. There are also studies in the literature that found that women are happier than men (Duffrin & Larsen, 2014; Ekici & Koydemir, 2013; Greenstein, 2016; Mookherjee, 1997; Lu, 2000; Wood, Rhodes, & Whelan, 1989).

While individuals who have never been married are more likely to be happy with success, job, money, and other sources of happiness than married individuals, they are less likely to be happy with health and love. This may be related to the fact that married individuals' motivation sources and priorities are their spouses or children. Thus, married individuals can care more about health and love. There are many studies in the literature stating that married individuals have a higher tendency to be happy (Akın & Şentürk, 2012; Atay, 2012; Bülbül & Giray, 2011; Ekici & Koydemir, 2013; Fernández-Ballesteros et al., 2001; Kangal, 2013; Çebi-Karaaslan et al., 2021; Lee, Seccombe, & Shehan, 1991; Myers 2000; Shinan-Altman, Levkovich, & Dror, 2020; Veenhoven & Dumludağ, 2015). On the contrary, there are studies that state that unmarried individuals have a higher tendency to be happy (Alexandre, Cordeiro, & Ramos, 2009; Kircı-Çevik & Korkmaz, 2014; Peterson et al., 2005).

As the education level of the individual increases, the probability of being happy with success increases. In the literature, Selim (2008) found that education has an important role in being happy with a job and money. This can be explained by the fact that educated individuals' achievements are more satisfying, especially when they do work related to their field. In addition, there are also studies that found the positive effects of the level of education on happiness (Atay, 2012; Bülbül & Giray, 2011; Chen & Short, 2008; Eren & Aşıcı, 2017; Kangal, 2013; Shinan-Altman et al., 2020) and the negative effects in the literature (Akın & Şentürk, 2012; Öndes, 2019; Servet, 2017).

An individual who is satisfied with his income level is more likely to be happy with his job, money, and other sources of happiness in his life. While an individual who is dissatisfied with his income level is less likely to be happy with success in life, the probability of being happy is higher with a job, money, and other sources of happiness. This situation may be related to the fact that success brings an improvement in the income level with it and that the individual who is not satisfied with the income level attaches importance to money and therefore to his job in order to improve it. In the literature, it is clear that income is one of the most basic factors affecting the happiness of individuals. There are many studies that found that individuals with financial independence are happier (Chen & Short) and that income has a positive effect on the happiness of individuals (Akın & Şentürk, 2012; Atay, 2012; Blanchflower & Oswald, 2004; Di Tella, MacCulloch, & Oswald, 2003; Diener & Diener, 2009; Ekici & Koydemir, 2013; Fernández-Ballesteros et al., 2001; Kircı-Çevik & Korkmaz, 2014; Veenhoven & Dumludağ, 2015).

Individuals who are satisfied with their lives are more likely to be happy with love than those who are less satisfied, and less likely to be happy with jobs, money, and other sources of happiness. In parallel with this result, individuals who are happy are less likely to be happy with success, job, money, and other sources of happiness, as well. This may be related to the achievement of spiritual satisfaction of these individuals. Likewise, an individual who is not satisfied with his social life is more likely to be happy with success. This situation may be related to the fact that individuals who are not satisfied with their social life keep their motivation areas in this direction by dedicating themselves to success in order to cover their deficiencies in that area of their lives. Social life is important for the happiness of individuals. In many studies in the literature, it has been determined that individuals who are satisfied with their social life and social relations are happier (Elliot, Cullen, & Calitz, 2018; Fernández-Ballesteros et al., 2001; Çebi-Karaaslan et al., 2021; Myers, 2000; Öndes, 2019; Sirgy & Cornwell, 2001). In addition, Chen & Short (2008) found that individuals living with their families were happier than those living alone.

The factors affecting the happiness and sources of happiness of individuals have had great importance from past to present. Being happy is among the most basic needs of individuals. Likewise, Maslow's hierarchy of needs states that the more an individual's needs are met, the happier the individual will be (as cited in Elliot et al., 2018).

In this study, important deductions were made about the factors affecting the happiness of individuals and their sources of happiness. The outputs obtained are presented in comparison with the literature, and attention is drawn to parallel and opposite situations. It has been hoped that the results of the study will shed light on the activities of policymakers and decision-makers who have an impact on individuals, or societies, experts working in this field.

REFERENCES

- Akın, H. B., & Şentürk, E. (2012). Bireylerin mutluluk düzeylerinin ordinal lojistik regresyon analizi ile incelenmesi. *Öneri Dergisi*, 10 (37), 183-193.
- Alkan, Ö., & Abar, H. (2020). Determination of factors influencing tobacco consumption in Turkey using categorical data analyses1. *Archives of Environmental & Occupational Health*, 75(1), 27-35.
- Alexandre, T. D. S., Cordeiro, R. C., & Ramos, L. R. (2009). Factors associated to quality of life in active elderly. *Revista de saude publica*, 43, 613-621.
- Atay, B. (2012). *Happiness in East Europe in comparison with Turkey*. (Yüksek lisans Tezi). İstanbul Bilgi Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.
- Ben-Akiva, M., & Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science* (pp. 5-33). Springer, Boston, MA.
- Blanchflower, D. G., & Oswald, A. J. (2004). Well-being over time in Britain and the USA. *Journal of public economics*, 88(7-8), 1359-1386.
- Bussi re, C., Sirven, N., & Tessier, P. (2021). Does ageing alter the contribution of health to subjective well-being?. *Social Science & Medicine*, 268, 113456-113465.
- B lb l, S., & Giray, S. (2011). Sosyodemografik  zellikler ile Mutluluk Algısı Arasındaki İlişki Yapısının Analizi. *Ege Academic Review*, 11(Special Issue), 113-123.
- Carandang, R. R., Shibanuma, A., Asis, E., Chavez, D. C., Tuliao, M. T., & Jimba, M. (2020). "Are Filipinos Aging Well?": Determinants of Subjective Well-Being among Senior Citizens of the Community-Based Engage Study. *International Journal of Environmental Research and Public Health*, 17(20), 7636-7649.
- Chen, F., & Short, S. E. (2008). Household context and subjective well-being among the oldest old in China. *Journal of family issues*, 29(10), 1379-1403.
- Çağlayan-Akay, E., & Timur, B. (2017). Kadınların ve erkeklerin mutluluğunu etkileyen fakt rlerin genelleştirilmiş sıralı logit modeli ile analizi. *Sosyal Bilimler Araştırma Dergisi*, 6(3), 88-105.
- Çebi-Karaaslan, K., Çalmaşur, G., & Emre-Aysin, M. (2021). Bireylerin Yaşam Memnuniyetlerini Etkileyen Fakt rlerin İncelenmesi. *Atat rk Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 35(1), 263-290.
- Di Tella, R., MacCulloch, R., & Oswald, A. (2003). The macroeconomics of happiness. *Review of Economics and Statistics*, 85, 809–827.
- Diener, E., & Diener, M. (2009). Cross-cultural correlates of life satisfaction and self-esteem. In *Culture and well-being* (pp. 71-91). Springer, Dordrecht.
- Diener, E. (2016). Happiness: The science of subjective well-being. *Noba textbook series: Psychology*. Champaign, IL: DEF publishers. DOI: <https://doi.org/https://doi.org/nobaproject.com>.
- Diener, E., Sapyta, J. J., & Suh, E. (1998). Subjective well-being is essential to well-being. *Psychological inquiry*, 9(1), 33-37.
- Diener, E., Suh, M. E., Lucas, E. R. ve Smith, H. (1999). "Subjective Well-Being: Three Decades of Progress", *Psychological Bulletin*, 125(2), 276–302.
- Duffrin, C., & Larsen, L. (2014). The effect of primary care fellowship training on career satisfaction, happiness and perceived stress. *Postgraduate medical journal*, 90(1065), 377-382.
- Ekici, T., & Koydemir, S. (2014). Social capital, government and democracy satisfaction, and happiness in Turkey: A comparison of surveys in 1999 and 2008. *Social Indicators Research*, 118(3), 1031-1053.
- Elliot, M., Cullen, M., & Calitz, A. P. (2018). Happiness among South African private sector physiotherapists. *The South African journal of physiotherapy*, 74(1), 1-10.
- Eren, K. A., & Aşıcı, A. A. (2017). The determinants of happiness in Turkey: Evidence from city-level data. *Journal of Happiness Studies*, 18(3), 647-669.
- Fernández-Ballesteros, R., Zamarr n, M. D., & Ruiz, M. A. (2001). The contribution of socio-demographic and psychosocial factors to life satisfaction. *Ageing & Society*, 2, 1-28.
- Garrow, L. A. (2016). *Discrete choice modelling and air travel demand: theory and applications*. England: Ashgate Publishing Limited.
- Greene, W. H. (2002). *Econometric analysis*. New Jersey: Pearson Education India.
- Greenstein, T. N. (2016). Gender, Marital Status, and Life Satisfaction: A Cross-National Study. In *Annual Meetings of the American Sociological Association, in Seattle, USA [United States of America], on August* (Vol. 21), 1-17.
- Kangal, A. (2013). Mutluluk  zerine kavramsal bir deęerlendirme ve T rk hanehalkı i in bazı sonu lar. *Elektronik Sosyal Bilimler Dergisi*, 12(44), 214–233.
- Kırcı-Çevik, N. K., & Korkmaz, O. (2014). T rkiye’de yaşam doyumunu ve iř doyumunu arasındaki ilişkinin iki deęişkenli sıralı probit model analizi. *Nięde Üniversitesi İktisadi ve İdari Bilimler Fak ltesi Dergisi*, 7(1), 126-145.

- Koppelman, F. S., & Wen, C. H. (1998). "Alternative nested logit models: structure, properties and estimation". *Transportation Research Part B: Methodological*, 32(5), 289-298.
- Larson, R. (1978). Thirty years of research on the subjective well-being of older Americans. *Journal of Gerontology*, 33(1), 109-125.
- Lee, G. R., Seccombe, K., & Shehan, C. L. (1991). Marital status and personal happiness: An analysis of trend data. *Journal of Marriage and the Family*, 3(4), 839-844.
- Lu, L. (2000). Gender and conjugal differences in happiness. *The Journal of social psychology*, 140(1), 132-141.
- Miñarro, S., Reyes-García, V., Aswani, S., Selim, S., Barrington-Leigh, C. P., & Galbraith, E. D. Happy without money: Minimally monetized societies can exhibit high subjective well-being. *PLOS ONE*, 16(1), 1-15.
- Mookherjee, H. N. (1997). Marital Status, Gender, and Perception of Well-Being. *The Journal of Social Psychology*, 137(1), 95-105.
- Moyano-Diaz, E., Mendoza-Llanos, R., & Paez-Rovira, D. (2021). Psychological well-being and their relationship with different referents and sources of happiness in Chile. *Revista de Psicología*, 39(1), 162-182.
- Muthuri, R. N. D. K., Senkubuge, F., & Hongoro, C. (2020). Determinants of happiness among healthcare professionals between 2009 and 2019: a systematic review. *Humanities and Social Sciences Communications*, 7(1), 1-14.
- Myers, D. G. (2000). The funds, friends, and faith of happy people. *American psychologist*, 55(1), 56-67.
- Newman, J. P., Lurkin, V., & Garrow, L. A. (2018). "Computational methods for estimating multinomial, nested, and cross-nested logit models that account for semi-aggregate data". *Journal of choice modelling*, 26, 28-40.
- Ngamaba, K. H., Panagioti, M., & Armitage, C. J. (2017). How strongly related are health status and subjective well-being? Systematic review and meta-analysis. *The European Journal of Public Health*, 27(5), 879-885.
- Öndes, H. (2019). Türkiye’de mutluluk düzeyini etkileyen faktörler: mekânsal ekonometri analizi. *Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 21(4), 1039-1064.
- Peterson, C., Park, N., & Seligman, M. E. (2005). Orientations to happiness and life satisfaction: The full life versus the empty life. *Journal of happiness studies*, 6(1), 25-41.
- Pollock, N. C., Noser, A. E., Holden, C. J., & Zeigler-Hill, V. (2015). Do Orientations to Happiness Mediate the Associations Between Personality Traits and Subjective Well-Being? *Journal of Happiness Studies*, 17(2), 713-729.
- Schnepf, S. V. (2010). Gender differences in subjective well-being in Central and Eastern Europe. *Journal of European Social Policy*, 20(1), 74-85.
- Scorsolini-Comin, F., & Santos, M. A. D. (2011). Relations between subjective well-being and marital satisfaction on the approach of positive psychology. *Psicologia: Reflexão e Crítica*, 24(4), 658-665.
- Selim, S. (2008). Türkiye’de bireysel mutluluk kaynağı olan değerler üzerine bir analiz: multinomial logit model. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 17(3), 345-358.
- Servet, O. (2017). Mutluluğun Türkiye’deki Belirleyenlerinin Zaman İçinde Değişimi. *Akdeniz İİBF Dergisi*, 17(35), 16-42.
- Shinan-Altman, S., Levkovich, I., & Dror, M. (2020). Are daily stressors associated with happiness in old age? The contribution of coping resources. *International Journal of Gerontology*, 14(4), 293-297.
- Sirgy, M. J., & Cornwell, T. (2001). Further validation of the Sirgy et al.'s measure of community quality of life. *Social Indicators Research*, 56(2), 125-143.
- Turkish Statistical Institute[TURKSTAT] (2021). Retrieved from <http://www.tuik.gov.tr>
- Veenhoven, R. & Dumludağ, D. (2015). İktisat ve mutluluk. *İktisat ve Toplum Dergisi*, 58(2), 46-51.
- Vijverberg, W. P. (2011). Testing for IIA with the Hausman-McFadden Test. IZA Discussion Papers 5826. *Institute for the Study of Labor (IZA)*, 1-52.
- Winkelmann, R. (2005). Subjective well-being and the family: Results from an ordered probit model with multiple random effects. *Empirical Economics*, 30(3), 749-761.
- Wood, W., Rhodes, N., & Whelan, M. (1989). Sex differences in positive well-being: A consideration of emotional style and marital status. *Psychological bulletin*, 106(2), 249-264.

Bireylerin Mutluluk Kaynaklarını Etkileyen Faktörlerin Multinomial Lojistik Modelle Analizi

Giriş

Bireylere, yaşamlarında kendileri için önemli olduğunu düşündükleri şeyler mutluluk getirir. Bu açıdan bakıldığında, her bireyin kendine özgü değerleri ve hedefleri vardır. Yani her bireyin mutluluk için farklı nedenleri vardır. Bu durum bireylerin mutluluk kaynaklarını etkileyen faktörlerin incelenmesini değerli kılmış ve bu çalışma için bir motivasyon kaynağı olmuştur. Çalışmanın amacı, bireyler, karar vericiler ve politika yapımcılar için bireylerin dahası toplumların ruhuna dokunacak, onları mutlu kılma noktasında katkı sağlayacak bir rehber olmaktır.

Çalışmada şu sorulara yanıt aranmaktadır: Demografik faktörler bireylerin mutluluk kaynakları üzerinde etkili midir? Ekonomik faktörler bireylerin mutluluk kaynakları üzerinde etkili midir? Bireysel faktörler bireylerin mutluluk kaynakları üzerinde etkili midir?

Bireylerin hayatında vazgeçilmez bir duygu olan mutluluk, hem birey hem de o bireyin oluşturduğu toplum için oldukça önemlidir. Bireylerin mutluluğu, mutlu toplumları beraberinde getirir, böylece toplumsal barış beslenir. Bu noktada mutluluk kavramı her bilim için oldukça önemlidir ve psikolojiden ekonomiye literatürde geniş yer bulmuştur. Ayrıca literatürde mutluluk kavramı, öznel iyi oluş ve yaşam memnuniyeti kavramları ile iç içe geçmiştir. Winkelmann (2005) çalışmasında, öznel iyi oluş ile aile arasındaki ilişkiyi incelenmiştir. Çalışma sonucunda; yaş ile öznel iyi oluş arasında “u” şeklinde ilişki olduğu, işsizliğin öznel iyi oluşu olumsuz etkilediği ve sağlığın öznel iyi oluşun önemli bir belirleyici olduğu tespit edilmiştir. Selim (2008) tarafından, mutluluk kaynağı değerleri analiz edilmiştir. Çalışmanın veri seti Türkiye İstatistik Kurumu aracılığıyla temin edilmiştir ve 6663 anket verisi ile çalışılmıştır. Çalışmada multinomial lojistik regresyon modeli kullanılmıştır. Çalışma sonucunda; gelirin mutluluk getirmediği, yaş arttıkça her mutluluk kaynağından olan tatmin seviyesinin düştüğü tespit edilmiştir. Bülbül ve Giray (2011) tarafından, sosyodemografik özellikler ile mutluluk algısı arasındaki ilişki araştırılmıştır. Çalışmada Türkiye İstatistik Kurumu tarafından yapılan 2008 yılı Yaşam Memnuniyeti Anket’i kullanılmıştır ve 6382 anket verisi ile çalışılmıştır. Çalışmada doğrusal olmayan kanonik regresyon analizi kullanılmıştır. Çalışmanın sonucunda bir işi olan, ortaokul mezunu ve geliri düşük olan erkeklerin mutluluk düzeyini orta ve üst olduğu, mutluluk kaynaklarının tüm aileleri olduğu; ilkokul mezunu, orta yaşlı, emeklilerin mutluluk düzeyinin orta ve üst olduğu tespit edilmiştir. Scorsolini-Comin ve Santos (2011) tarafından, evlilik ile öznel iyi oluş arasındaki ilişki incelenmiştir. Çalışmaya 53 çift katılmıştır. Çalışmada veri setinin analizi için korelasyon ve çoklu regresyon analizleri yapılmıştır. Bireylerin yaş ortalaması 42’dir. Çalışma sonucunda öznel iyi oluşun evlilik durumu üzerinde olumlu etkisinin olduğu tespit edilmiştir. Akın ve Şentürk (2012) tarafından, bireylerin mutluluk düzeyini etkileyen değişkenler incelenmiştir. Çalışmada, 2007 yılı Avrupa Yaşam Kalitesi Anketi kullanılmıştır ve sıralı lojistik regresyon analizi uygulanmıştır. Çalışma neticesinde; mutluluk düzeyinin demografik özellikler açısından farklılaşmasına rağmen temelde benzer sonuçlar verdiği, yaşın eğitim seviyesinin artışıyla mutluluğun azaldığı, erkeklerin kadınlara kıyasla daha mutlu olduğu, evli ve sağlıklı olmanın mutluluğu olumlu etkilediği tespit edilmiştir. Çağlayan-Akay ve Timur (2017) tarafından, kadınlar ve erkeklerin mutluluğu üzerinde etkili olan faktörler araştırılmıştır. Çalışmanın veri seti Türkiye İstatistik Kurumu aracılığıyla temin edilmiştir ve çalışmada genelleştirilmiş sıralı lojistik regresyon modeli kullanılmıştır. Çalışma sonucunda; ekonomik faktörlerin mutluluk üzerinde etkili olduğu, yaşın mutlu olma olasılığını arttırdığı, iş yerinde çalışmanın ve iş yeri açmanın mutlu olma üzerinde olumlu etkisinin olduğu, umutlu olmanın kadınlar ve erkekler için mutlu olma olasılığını artırıcı olduğu tespit edilmiştir. Shinan-Altman, Levkovich ve Dror (2020) tarafından, yaşlı bireylerin mutlulukları üzerine bir araştırma yapılmıştır. Çalışma veri seti İsrail ‘de anket uygulaması aracılığıyla toplanmıştır ve verilerin analizi için hiyerarşik regresyon analizi uygulanmıştır. Çalışma sonucunda; bireylerin mutluluk düzeylerinin orta düzeyli olduğu, iyimserlik ve sosyal desteğin mutluluğu olumlu etkilediği, evlilerin bekarlara kıyasla daha mutlu olduğu, eğitim ve gelirin mutluluk

üzerinde olumlu etkisinin olduğu, cinsiyet ve yaşın mutluluk üzerinde anlamlı bir etkisinin olmadığı tespit edilmiştir. Bussière, Sirven ve Tessier (2021) tarafından sağlık ile öznel iyi oluş arasındaki ilişki araştırılmıştır. Çalışmanın veri seti on Avrupa ülkesini içeren bir anket uygulaması aracılığıyla elde edilmiştir. Çalışma sonucunda sağlığa verilen önemin zamanla farklılaştığı, yaşlanmanın bireyler için sağlığın öznel refah üstündeki etkisini arttırdığı ve sağlık ile öznel refah arasındaki ilişkiyi güçlendirdiği tespit edilmiştir. Minarro vd. (2021) tarafından, para ile öznel iyi oluş arasındaki ilişki incelenmiştir. Çalışmanın veri seti Solomon Adaları ve Bangladeş'teki kıyı topluluklarına anket yapılarak elde edilmiştir. Çalışma sonucunda, ekonomik büyümenin düşük gelirli topluluklarda yaşam memnuniyeti arttırmayacağı, öznel iyi oluşun çok para kazanmayla elde edilemeyeceği tespit edilmiştir.

Bireylerin mutluluğu, yaşam memnuniyetleri ve öznel iyi oluşları üzerinde birçok faktör etkilidir. Demografik ve ekonomik faktörler literatürde en çarpıcı ve en yaygın olanlardır. Bireylerin yaşı, cinsiyeti, medeni durumu, eğitimi ve geliri birçok çalışmada karşımıza çıkmaktadır. Çalışmaların çoğunda, bu faktörler mutluluk düzeyi, yaşam memnuniyeti ve öznel iyi oluş üzerinde istatistiksel olarak anlamlı etkiler göstermiştir. Genel olarak yapılan araştırmalarda kadınların, evlilerin, eğitimlilerin ve geliri yüksek olanların daha mutlu olduğu tespit edilmiştir.

Yöntem

Çalışmada Türkiye İstatistik Kurumu tarafından yapılan Yaşam Memnuniyeti Anketi kullanılmış ve çalışmaya 9212 kişi dahil edilmiştir. Çalışmada veri düzenleme için Microsoft Excel, ki-kare analizleri için SPSS 20, multinominal lojistik regresyon analizi için Stata 14.1 programları kullanılmıştır. Öncelikle araştırmaya katılan bireyin mutluluk kaynağına göre frekans analizleri yapılmıştır. Bireylerin mutluluk kaynağı ile bağımsız değişkenler arasındaki ilişkiyi incelemek için ki-kare bağımsızlık testleri yapılmıştır. Daha sonra multinominal lojistik regresyon analizi kullanılarak bireylerin mutluluk kaynağına etki eden faktörler ve bu faktörlerin etki büyüklükleri belirlenmiştir.

Sonuç ve Tartışma

Çalışmadan elde edilen bulgulara göre; bireylerin %21,5'inin 38-47 yaş aralığında ve %54,1'i kadındır. Çalışmaya dahil edilen bireylerin büyük çoğunluğu (%72,8) evlidir. Bireylerin %13,7'si bir okul bitirmemişken %19,1'i üniversite mezunudur ve %57,8'i bir işte çalışmamaktadır. Bireylerin %40,8'i hanenin aylık gelirinden memnun ve çok memnun iken %41,1'inin refah düzeyi ortalamanın altındadır. Bireylerin %53,8'inin mutlu ve çok mutlu olduğu, %75,1'inin tüm aile bireyleri tarafından mutlu edildiği, %47,3'ünün yaşamından memnun olduğu, %67'sinin sağlığından memnun ve çok memnun olduğu, %54,9'unun aldığı eğitimden memnun ve çok memnun olduğu, %48'inin sosyal hayatından memnun ve çok memnun olduğu, %70,4'ünün kendi geleceğinden umutlu olduğu, %41,5'inin 5 yıl öncesi ile karşılaştırıldığında maddi manevi bugünkü durumunun gerilediği, %31,6'sının gelecek 5 yıllık dönem için genel olarak durumunun aynı kalacağını ifade ettikleri tespit edilmiştir.

Çalışma sonucunda yaş, cinsiyet, medeni durum, eğitim durumu, gelir düzeyinden memnuniyet, refah düzeyi, yaşam memnuniyeti, sosyal hayattan memnuniyet faktörlerinin bireylerin mutluluk kaynakları üzerinde etkili olduğu tespit edilmiştir. Çalışma aracılığıyla; koronavirüs salgınının psikoloji başta olmak üzere hayatımızın pek çok yönünü olumsuz etkilediğinin ve yarınlarımıza iz bırakacağına aşikâr olduğu böyle bir zamanda bireylerin mutlulukları arttırmak ve yarınların daha güzel olmasını sağlamak için karar vericilerin ve politika yapıcıların faaliyetlerine ışık tutulur.

Monitoring Student Achievement with Cognitive Diagnosis Model

Levent YAKAR*

Nuri DOĞAN **

Şenol DOST***

Nazan SEZEN YÜKSEL****

Abstract

In this study, it is aimed to show how student achievement can be monitored by using the cognitive diagnosis models. For this purpose, responses of the 6th, 7th, and 8th grade Mathematics subtests of High School Placement Tests (HSPT) in 2009, 2010, and 2011, which provide longitudinal data, were used, respectively. There were 49933 examiners' responses in data sets. The attributes examined by these tests were determined by the Mathematics experts, and the Q matrix consisting of five attributes was developed. As a result of the analysis, it was seen that the largest latent class for all three years consisted of those non-master for any attribute. It was observed that the probability of attribute mastery increased in the 7th grade and decreased in the 8th grade. The high classification accuracy seen as a result of the analysis applied to HSPT, which is not intended for the cognitive diagnosis, shows that the results can be used for monitoring student achievement.

Key Words: Cognitive diagnosis, student achievement, g-dina, attribute mastery probability, longitudinal data.

INTRODUCTION

Education includes the efforts made to gain individuals the pre-determined and necessary behaviors related to the cognitive, affective, and psychomotor areas. Gaining targeted behaviors are not operationsthat happen at once, but require a process. It can be said that this situation is also reflected in measurement and evaluation. Although, in measurement and evaluation practices, it is very common to collect data on the extent to which the product reached at the end of the process meets the expected qualifications, contemporary educational approaches accept that products are not independent of the processes and interactions in the process (Kutlu, Doğan & Karakaya, 2010). Therefore, it is necessary to measure the processes and interactions in the training process as well as the products at the end of the training process.

It is observed that as the importance is given to revealing the development of individuals in the process, practices and researches aimed at this purpose increase. If it is accepted that measurement practices related to the process are generally for monitoring the development, it can be said that the studies for gathering information about the process are carried out through both national and international exams (Ministry of National Education [MoNE], 2017; Organisation for Economic Co-operation and Development [OECD], 2019). For example, through international exams such as PISA, TIMSS, and PIRLS, national-level development is tried to be monitored by making use of international comparisons in areas such as mathematics, science and technology, and reading comprehension. Although international exams give indirect information about the educational process in general, they do not provide information about the status of individual students who are the subjects of the process and cannot provide detailed information about the effectiveness of educational programs. In this regard, it is observed that in many countries, information about the process is collected through exams held at

* Assist. Prof. Dr., Kahramanmaraş Sütçü İmam Uni, Faculty of Education, Kahramanmaraş-Turkey, l_yakar@hotmail.com, ORCID ID: 0000-0001-7856-6926

** Prof. Dr., Hacettepe Uni, Faculty of Education, Ankara-Turkey, nuridogan2004@gmail.com, ORCID ID: 0000-0001-6274-2016

*** Prof. Dr., Hacettepe Uni, Faculty of Education, Ankara-Turkey, dost@hacettepe.edu.tr, ORCID ID: 0000-0002-5762-8056

**** Assoc. Prof. Dr., Hacettepe Uni, Faculty of Education, Ankara-Turkey, nsezen.hc@gmail.com, ORCID ID: 0000-0002-0539-3785

To cite this article:

Yakar, L., Doğan, N., Dost, Ş., Sezen Yüksel, N. (2021). Monitoring student achievement with cognitive diagnosis model. *Journal of Measurement and Evaluation in Education and Psychology*, 12(3), 303-320. doi: 10.21031/epod.903084

Received: 25.03.2021
Accepted: 24.09.2021

different stages of education (OECD, 2019). For example, in the past Detection Exam of Student Achievement and today Monitoring and Evaluating Academic Skills exams aim to reveal the developments in main courses in Turkey (MoNE, 2017). In addition to these, although the past High School Placement Exam (HSPT; “Seviye Belirleme Sınavı”), which students took three times in secondary school, is not an application for monitoring, it can be said that it is a test that provides information in terms of student development due to its multiple implementations (MoNE, 2008).

It is seen that research models based on repeated measurements come to the fore in order to determine the development of individuals in the process. In order to reveal whether the development of individuals is sufficient or not in research models based on repetitive measurements, the measurement results made at least two different times are compared using various statistical techniques. However, since the results obtained from such applications are based on the comparison of the average of the measurement results obtained at two different times at least, it does not give information about the development of individual students as well as neglecting the acquisitions and subject dimension. These techniques are criticized in this respect (Lohman, 1999).

In addition to traditional statistical techniques, cognitive diagnosis models (CDM), which is an effective technique to reveal the fine-grained ability parameters of individuals, can demonstrate level developments in repeated measures. It is stated that CDM, which will be discussed in this study, has become widespread, especially with the beginning of the 2000s, and its main purpose is to give cognitive feedback to teachers and families about their students (Embretson, 1998).

CDM is based on latent class analyses, which are used to identify subgroups and determine which individuals belong to these subgroups using multivariate categorical data and interrelated situations (Cheng, 2010). In this way, it is possible to calculate the structure of certain knowledge or the development of a skill in the student by taking into account the strengths and weaknesses of the student at the cognitive level (Leighton & Gierl, 2007). According to de la Torre (2009), with a test developed using CDM, it can be determined which skills the students have, which are predetermined by experts, and therefore, what their shortcomings are. Taking advantage of this feature of CDM, it may be possible to see the development of students in terms of relevant skills if the same skills are measured at different times.

By using CDM, psychological structures with more than one interrelated cognitive attributes can be measured with a single test. In practice, it is accepted that each item in the test measures one or more cognitive attributes. In CDM analyses, the Q matrix is used to determine which item measures which cognitive attribute. In the Q matrix, each column represents a cognitive attribute, and each row represents an item. The Q matrix is created by field experts by coding as 1 if the cognitive attributes specified in the column are measured with the item specified in the row, and 0 if not (de la Torre & Minchen, 2014). By the Q matrix used in CDM, 2^k latent classes are formed for k cognitive attributes defined by experts. There will be eight latent classes for k = 3; the latent class (000), indicating an individual who is non-master for any attributes; latent classes (100), (010), (001) indicating individuals with master one of the attributes; latent classes (110), (101), (011) indicating individuals with master two of the attributes and (111) latent class indicating individuals with master all the attributes. In addition to showing what attributes individuals have and which they do not, the latent classes also give an idea of which questions they are expected to answer correctly. CDM makes it possible to identify individuals in terms of cognitive attributes.

There are many CDM available; Deterministic inputs noisy and-gate (DINA; Junker and Sijtsma, 2001), Deterministic inputs noisy or-gate (DINO; Templin & Henson, 2006), re-parameterized unified model (R-RUM; Hartz, 2002), general diagnostic models (GDM; von Davier, 2008), generalized DINA Model (G-DINA; de la Torre, 2011), etc., that take different assumptions and parameters into account. Besides the various test and item parameters, the mastery probability of cognitive attributes in the Q matrix is calculated to determine which of the latent classes individuals will be included in, in these models. If the probability values calculated for a cognitive attribute are 0.5 and above, mastery of attribute is shown with "1"; if it is less than 0.5, it is indicated with "0". This process aims to make it easier to reveal the latent cognitive structures that individuals have.

Considering the example given above, it can be said that individuals in the "101" latent class have the first and third cognitive attributes and their probability of mastery of these attributes is 0.5 or above. On the other hand, it can be said that these individuals do not have the second cognitive attribute and their probability of mastery of the second cognitive attribute is less than 0.5. Therefore, while latent classes are obtained by rounding the probability value to 0 or 1, the differences between the probabilities of individuals are neglected. For example, an individual who has mastery probability of the first, second, and third cognitive attributes, respectively 0.55, 0.10 and 0.60; and an individual who has probability mastery 0.90, 0.45, and 0.95 are in the same latent class, which coded with "101". The fact that the transformation of mastery of attribute probability into binary category causes loss of information can be seen as the negative side of this transformation process.

It is one of the most important features of CDMs that they reveal the attributes they have in smaller parts instead of the holistic approach when diagnosing individuals. In this way, CDMs enable individuals to be diagnosed from different angles. The latent classes and attribute mastery probability outputs that are created with the help of the Q matrix input representing fine-grained small measurement units in CDMs provide detailed information for individuals. The fact that monitoring the cognitive characteristics of students in fine-grained skill with CDMs can provide more specific and relevant information compared to the general monitoring of students' cognitive level reveals that CDMs will be more useful in monitoring students' progress.

Considering that the main purpose of CDMs is to provide feedback to education stakeholders (Embretson, 1998), a detailed and fine-grained picture of the current situation can be taken through CDMs. Formative assessment, in which feedback is at the forefront, cannot be used adequately due to the high class size, the need for time and effort (Bennett, 2011). In this case, it is important to include high-stakes exams, which are not normally intended for formative assessment, in the feedback mechanism. In addition, the longitudinal feedback to be given for the same parts with the same method will be of great importance in terms of revealing the change and making the education even better.

Interest in CDMs is increasing both in the world and in Turkey. It can be said that the field of study of CDMs is mostly focused on simulation since the subject area is new with increasing interest. In some of these studies (Huang, 2017; Kaya & Leite, 2017; Wang, Yang, Culpepper & Douglas, 2018; Zhan, Jiao, Liao & Li, 2019), models are presented for the use of longitudinal data in CDMs. However, these studies are insufficient to show how the change in a large population is revealed by CDMs. The actual data in these studies consist of smaller datasets suitable for model use only. This study, on the other hand, is important by separating it from other studies in terms of targeting a wide audience.

In this study, it was aimed to apply cognitive diagnosis models to HSPT, which are repeated measures, and to monitor the development of students through their attributes. For this purpose, answers to the following sub-problems were sought;

- 1) What is the prevalence rate of the latent class patterns of students by years?
- 2) What is the rate of change in the students' mastery of each attribute by years?
- 3) What is the rate of change in the number of attributes mastered by students over the years?
- 4) What is the level of reliability and validity of the findings obtained?

METHOD

In this study, which aims to monitor the achievement of students with CDM, survey method, one of the quantitative research methods, was used.

Sample

The population of the study consists of approximately 1 million middle school students who started secondary school in 2009 and joined HSPT in 3 years. The answers of 131068 of these students in the

SBS every three years were given to the researchers by the MoNE. The data of 49933 students, who had complete data in all three years, formed the sample of the research.

Data Collection Instruments

High Schools Placement Exams (HSPT) was organized by the Ministry of National Education (MoNE), Turkey. The data were obtained from the General Directorate of Measurement, Evaluation, and Examination Services of the MoNE upon the request of the researchers. HSPT was a central system high-stake exam which was held after the course period in every year in June, organized by the MoNE, and measures the level of achievement of the students related to the learning outcomes determined for the 6th, 7th, and 8th grades. The exam consisted of Turkish language, mathematics, science, social sciences, and English courses items within the scope of the middle school. The exam was prepared as multiple-choice tests based on the learning outcomes and is sufficient to measure the student's interpretation, analysis, critical thinking, predicting, and problem-solving skills, etc. (MoNE, 2010).

HSPT was an exam held once a year between 2008 and 2013 at the end of the spring term, and its scores are used to place students in high school. Approximately 1 million students had entered HSPT for each grade level each year. HSPT differs from the other old/new exams in terms of being held in 6th, 7th, and 8th grades among the exams held for transition to high schools. With this feature, HSPT is an important resource to examine the development of students over the years. Although the exam is not practiced today, HSPT was deemed suitable for this study because it has been measured more than twice, the answers of many students across the country have been obtained, and the study is on a theoretical and practical basis.

In the study, in which student progress was examined through math test items, 16, 18, and 20 math items were asked to students in the 6th, 7th and 8th grades, respectively.

Descriptive statistics regarding the test scores of the data used in the study are calculated and given in Table 1.

Table 1. Descriptive Statistics for Tests

Grade	Item	Mean	Std. Dev.	Mean Item Difficulty	<i>d</i>
6 th	16	5.40	2.93	0.338	-.82
7 th	18	7.62	4.44	0.423	-.4
8 th	20	6.65	5.10	0.333	-.97

Table 1 shows that it was observed that the highest number of correct answers was in the 7th grade, while the lowest was in the 6th grade. It was seen that 8th grade students had correct responses on average 1.25 more questions than 6th grade. However, when the mean item difficulty, which indicates rates of correct responses, are examined, it is seen that the 6th and 8th grades are very close to each other due to the increasing number of questions over the years. Considering the relative variation coefficients showing the ratio of the standard deviation to the mean, it can be said that the groups become more heterogeneous from the sixth to the eighth grade. When the difficulty level of the tests is evaluated, the tests applied for the 6th and 8th grades have a similar difficulty, and the tests applied to the seventh grade are relatively easy. In the analysis for the item response theory, it was seen that all three data sets were two-dimensional. When the averages of the item difficulty parameters (*d*) obtained as a result of the analysis are examined, it can be said that the items in the 8th grade are easier. Although the 7th grade items are a little more difficult than the 8th grade, it is concluded that they are easy. It was seen that the 6th grade items were more difficult than the other grades but still close to the easy level.

Procedure

Defining Attributes

In the primary mathematics teaching program, problem-solving ability is one of the basic skills that are stipulated to be provided to the students. Within the scope of the program, the problems are discussed under two headings as routine and non-routine. In general, problems are considered as questions, of which solutions are not foreknown and obvious, and in such questions, it is claimed that the students will reach a solution by making reasoning through their current knowledge (Sezen Yüksel, Sağlam Kaya, Urhan, & Şefik, 2019). In brief, problems that can be solved by using the information directly are described as "routine" problems, whereas problems that can be solved by interpreting existing information and by operations that are more complex are described as "non-routine problems".

Within the scope of this study, it was tried to determine the attributes of HSPT math items. For this purpose, the 6th, 7th, and 8th grade items were discussed primarily within the context of the problem types and then the mathematical skills that they require. While determining these skills, first of all, Math Taxonomy (Smith, Wood, Coupland, Stephenson, Crawford, & Ball, 1996) and "Mathematical content and process skills" (Tatsuoka, Corter, & Tatsuoka, 2004) in the literature were examined, and the operations required by the HSPT mathematics questions were grouped by the field expert researchers of the study. Operations (such as the application of a well-known algorithm, visual perception) that could not be classified into any of the existing skills were determined by field experts, gathered under common categories, and renamed. Five attributes have been created by making these skills more specific in accordance with the processes and subjects required by the items. The names and characteristics of these attributes are as follows:

Table 2. Defined Attributes' Code, Name and Definitions

Attribute Code	Attribute Name	Definition
Attribute 1	Operational Applications	Application of Basic Features of Numbers
Attribute 2	Mathematization Applications	Mathematization of a Word Problem
Attribute 3	Concept Calculations	Computational Application of Concept
Attribute 4	Concept's Advanced Applications	Application of the concept, in a different context in relation to other concepts, in a problem
Attribute 5	Geometric Manipulation	Application for Manipulation of Geometric Shapes

Attribute 1 covers the processes of "Routine operations by recalling a definition or a term, application of the formula, recalling the rules knowledge, classification knowledge, applying an algorithm, length measurement, numbers (fractions, decimal numbers, and percentages) and determining the number representations and making number conversions", which includes application of the basic features of the numbers.

Attribute 2 covers the mathematization of a word problem. In other words, it is the process of problem-solving through the use of mathematical representations of verbal expressions containing mathematical structures and taken from daily life.

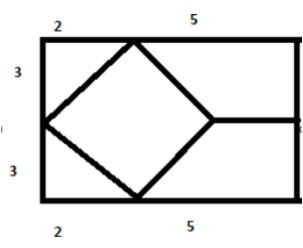
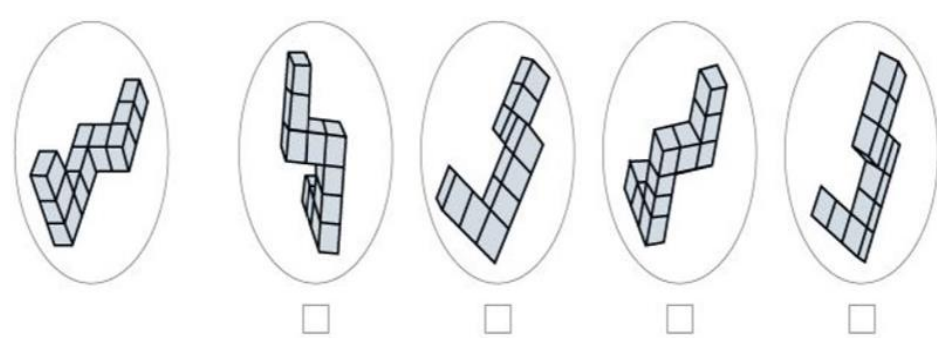
Attribute 3 includes the processes regarding the computational applications of concepts. This stage is the application of the processes required by the mathematical concept given in the problem expression.

Attribute 4 is operations of applying concepts in a different context and associated with other concepts.

Attribute 5 covers applications for the manipulation of geometric shapes. These applications involve the use of different forms of geometric shapes provided in the problem.

It would be beneficial to examine Table 3 to make the information on the attributes more understandable:

Table 3. Example Items Which Examined Attributes

Attribute No	Example Items
Attribute 1	What is the result of the operation $(-2)^{-3} \cdot 4^2$?
Attribute 2	“Ayşe has 440 pieces 1 TL coins in her penny bank. Ayşe spent all her money to buy 5 dolls. In that case, what is the price of a doll?”
Attribute 3	How many unit squares is the area of a circumscribed circle of the square with a side length of 4 cm?
Attribute 4	 <p>2 triangular, 2 trapezoidal, and 1 equilateral rectangular regions are drawn in the rectangular region of the figure. When Ela throws a stone, what is the probability of the stone striking the triangular regions given in the figure?</p>
Attribute 5	 <p>Which one of the figures given in the above <u>cannot be obtained</u> by rotating the leftmost shape?</p>

Five field experts were consulted for the mathematical attributes determined by the researchers. The field experts consisted of two academicians with specialisation in mathematics education and three mathematics teachers who were working in schools affiliated to the Ministry of National Education at the time of the study. Initially, their opinions about the names and contents of the attributes were elicited. The definitions and content of some attributes were modified based on these opinions. For example, due to the fact that the skills of the application of the basic features of the numbers given in the content of the Attribute 1 were perceived as four operations at first glance, an error was identified as considering that this attribute was included in all problems. In order to eliminate this error, it was decided to use more specific concepts in the definition of Attribute 1. Therefore, Attribute 1 was expressed as applications related to the basic characteristics defined on the number sets. Another correction suggestion encountered at this stage was related to the items of geometric shapes. Geometry has its own specific framework, and it is possible to solve some questions by known algorithms as such in mathematical questions. In accordance with the feedback taken from the experts, Attribute 5 was renamed as “Geometric Manipulation” because it did not address all geometry questions because the skills required by the solution of some geometry questions were the applications of the known algorithm. This led to the tagging of geometry items with other attributes, although the word geometry was not used in the attribute. Subsequently, the revised mathematical attributes were re-shared with the field experts in concern. In consequence, a consensus was reached on the attributes in accordance with the opinions received, and the attributes and their explanations were finalized accordingly.

Creating of Q-matrix

In general, the problems in mathematics differ from each other in the context of the mathematical skills required by their content and solution. This was taken into account when tagging the items according to the attributes established within the scope of the study. The lack of a hierarchical structure among mathematical skills leads to the lack of a hierarchy between the mathematical attributes prepared according to these skills. These facts played a significant role in the formation of the Q-matrix. For instance, a problem tagged as Attribute 4 may not contain other attributes. On the other hand, an item can be tagged with more than one attribute. For example, we may consider the problem of "Each one of the T-shirts purchased for TL 4,50 is printed on TL 1,25. When these t-shirts are sold to TL 9,50, which of the following is the algebraic expression of the profit earned from x unit?". This problem is tagged with Attributes of 1-2-3 because it includes basic operations with decimal numbers, mathematization of a word problem, and computational application of the concept of "profit". The four operations used on any problem are not always required to refer to Attribute 1, though.

In the process of obtaining the Q-matrix, the researchers firstly formed matrices individually. Then, they came together to compare the matrices. In this process, when the items were tagged with different attributes, the researchers finalized the matrix by reaching a consensus by revising the mathematical skills included in the questions.

For the Q-matrix formed in the last case, the opinions of two academicians from the field experts who took part in the beginning of the process were obtained. The Q-matrix and items were submitted to the field experts together with the explanations of the attributes. The suggestions taken from both field experts were evaluated together. In order to give an example of the correction suggestions, item 19 of 8th graders' HSPT can be examined. In this item, a ramp image and the height of this ramp were given as 1 meter, and the slope was 10%, and if the slope was 8%, it was asked what point the ramp would start from the visual point. The researchers tagged this question with attributes 1, 4, and 5. The feedback received from the field expert was that this question did not include a geometric shape; therefore, it would not be related to Attribute 5. The researchers emphasized that the ramp image contained in this question covered a right triangle and that a solution could be reached through its manipulation. The field experts reached a consensus on this issue. A similar method was followed for other suggestions, and the Q-matrix was finalized by consensus with the field experts.

Some modification suggestions based on the results of the data-model fit of Q matrices, AIC (Akaike, 1974), BIC (Schwarz, 1976), and the software package developed for Q matrix validation (Ma ve de la Torre, 2019) were conveyed to experts. The relative fit indices before and after the last recommendation are presented in Table 4.

Table 4. Relative Fit Indices Before and After the Last Recommendation

	AIC		BIC	
	Previous	Last	Previous	Last
6	923572.8	923036.4	924322.4	923741.9
7	1013470	1012564	1014114	1013225
8	1018767	1014285	1019499	1014999

As seen in Table 4, AIC and BIC relative model data fit indices at all three grade levels indicate a better fit for the Q matrices formed after the accepted recommendations. In line with the analyses and suggestions, the Q matrices were given their final form in Table 5.

According to Table 5, in the last case, Attribute 1 was examined in ten items in the 6th grade, four items in the 7th grade, and five items in the 8th grade. Attribute 2 was examined in six items in the 6th grade, three items in the 7th grade, and four items in the 8th grade. Attribute 3 was examined in seven items in the 6th grade, seven items in the 7th grade, and eight items in the 8th grade. Attribute 4 was examined in three items in the 6th grade, three items in the 7th grade, and four items in the 8th grade. Attribute 5 was examined in two items in the 6th grade, five items in the 7th grade, and six items in the 8th grade. In six of the 6th grade items, one attribute was examined, in nine of them two, and in one of them, four

were examined. In 14 of the 7th grade items, one attribute was examined, and in four of them, two attributes were examined. In 13 of the 8th grade items, only one attribute and in seven of them, two attributes were examined.

Table 5. Q Matrix

6 th Grade						7 th Grade					8 th Grade						
Item	A1	A2	A3	A4	A5	Item	A1	A2	A3	A4	A5	Item	A1	A2	A3	A4	A5
1	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0
2	1	0	1	0	0	2	1	0	0	0	0	2	1	1	0	0	0
3	1	1	0	0	0	3	0	0	0	1	1	3	0	0	1	0	0
4	1	1	0	0	0	4	0	0	1	0	0	4	0	0	1	0	0
5	1	0	1	0	0	5	1	1	0	0	0	5	1	0	0	1	0
6	0	0	1	0	0	6	0	0	0	0	1	6	0	0	1	0	0
7	0	0	1	0	0	7	0	0	1	0	0	7	1	1	0	0	0
8	1	0	1	0	0	8	0	0	0	1	0	8	0	0	1	0	1
9	0	0	0	0	1	9	0	1	0	0	0	9	0	0	0	0	1
10	1	1	1	1	0	10	0	0	1	1	0	10	0	0	0	0	1
11	0	0	0	1	0	11	0	0	0	0	1	11	0	1	0	0	0
12	1	1	0	0	0	12	0	0	1	0	0	12	0	0	1	0	1
13	0	0	0	0	1	13	0	0	1	0	0	13	0	0	1	0	1
14	1	1	0	0	0	14	0	0	1	0	0	14	0	0	1	0	0
15	1	0	1	0	0	15	0	0	0	0	1	15	0	0	0	1	0
16	1	0	0	1	0	16	0	0	1	0	0	16	0	1	0	0	0
						17	0	1	0	0	0	17	0	0	0	1	0
						18	1	0	0	0	1	18	0	0	1	0	0
												19	1	0	0	1	0
												20	0	0	0	0	1

Data Analysis

In the selection of the model to be used in order to determine the cognitive classes of the students, the criterion of the model having the best fit at the item level was taken into consideration. For this purpose, the data sets and the Q matrices they were related to were subjected to model comparison analysis with the GDINA (Ma and de la Torre, 2018) package in the R software program. It was tested with Wald test that shows which of the G-DINA in the package or the restricted forms of G-DINA, DINA, DINO, ACDM, LLM (Maris, 1999), R-RUM (DiBello, Stout, & Liu Roussos, 2007) fit better. If there was no significant difference at the $p=.05$ level between the fit indices of G-DINA and its restricted forms, the restricted model with the simplest structure was chosen; otherwise, G-DINA was chosen as the model to be used for the relevant item. As a result of the analysis, in the 6th grade, LLM for the 3rd item, DINA for the 4th item, and R-RUM for the 10th and 12th items were determined as the most appropriate model. And in 8th grade, LLM for the 7th item and the R-RUM model for the 8th, 12th, and 19th items were determined as the most appropriate model. The GDINA model was determined as the most appropriate model for all the items in the 7th grade and for the other items in the 6th and 8th grades.

Analyses were performed using the R software program using the GDINA (Ma and de la Torre, 2018) package. Expected a Posteriori (EAP) method was used to obtain individual parameters. For the first research question, the probability of mastering each attribute by years and the prevalence rates of the latent classes to which they were assigned as a result of the analysis were given. For the second sub-problem, the rates of change according to the years of mastery of each attribute are given. In the third sub-problem, the changes in the number of attributes of the students according to the years were reported. In the last sub-problem, the correct classification rates were examined for the reliability of the analysis results (Ciu, Gierl, Chang, 2012). For this, the accuracy of latent classifications was determined by Iaconangelo (2017), and the accuracy of classification by attribute was determined by Wang et al. (2015) with the help of indexes in the same package. In the examination of the validity of the analysis results, the proof of convergent validity was used. The correlation between the correct response rate and the probability of mastery of the attribute was examined as proof of convergent validity (Li, et al., 2020).

RESULTS

Findings are given in order under sub-headings according to the sub-problem titles.

Students' Attributes by Years

The average of the students' attribute mastery probability for each year (grade level) was calculated and given in Table 6.

Table 6. Means of Attribute Mastery Probability

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Mean
6 th Grade	0.22	0.17	0.13	0.12	0.16	0.16
7 th Grade	0.23	0.31	0.22	0.17	0.37	0.26
8 th Grade	0.17	0.27	0.16	0.34	0.25	0.24

According to Table 6, it is seen that all attribute mastery probability increased with the transition from 6th to 7th grade. The probabilities of all the 8th grade attributes except attribute 1 were also found higher than the 6th grade levels. When the 8th grade probabilities were compared with those of the 7th grade, the values revealed closer to each other, but it is seen that the probabilities in the 7th grade were higher for all the qualities except attribute 4.

In the sixth grade, the most common attribute was attribute 1, followed by the 2nd and 5th attributes with similar rates. The least attribute mastery probability in the sixth grade was observed in attributes 4 and 3. In addition, the probability value of attribute 4 in the sixth grade was seen to have the lowest value among all the attribute probabilities covering three years. The highest attribute mastery probability in the seventh grade was in attribute 5, followed by attribute 2. In the 7th grade, attribute 4 had the lowest probability. The highest attribute mastery probability in the 8th grade was in attribute 4. In the 8th grade, it is seen that attributes 1 and 3 had the lowest probability average. The attribute mastery probability and the correct response rate for each class are given in Figure 1.

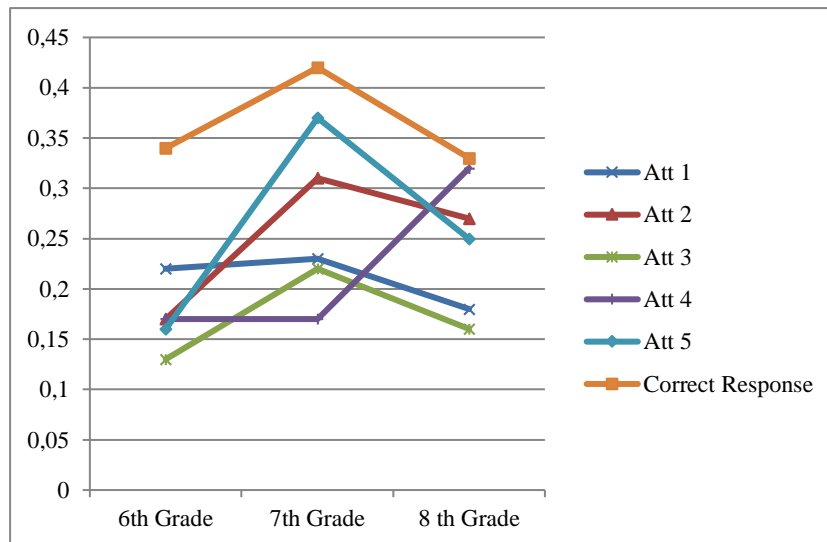


Figure 1. Correct Response Rate and Attribute Mastery Probability Means across Years

Figure 1 shows the variation of attribute mastery probability and correct response rate and their relation by years, clearly. It can be concluded that while the correct response rate and the mastery probability of attributes 2, 3, 4, and 5 increase visibly, the mastery probability of attribute 1 did not change much at the time of the transition from the 6th grade to the 7th grade. When transitioning from the 7th grade to the

8th grade, a decrease was observed in the rate of correct response and all attribute mastery probability except attribute 4. Among the attributes, attribute 4 had the lowest probability average in the 6th and 7th grades and the highest probability average in the 8th grade.

The average probability of having attributes may not provide sufficient information about the attribute patterns of students. For this, the latent attribute classes were examined. Table 7 shows the rate of students in latent attribute classes for 3 years.

Table 7. Prevalence of Latent Classes

Latent Class	6 th Grade	7 th Grade	8 th Grade	Latent Class	6 th Grade	7 th Grade	8 th Grade
00000	0.75	0.62	0.67	11100	0.00	0.00	0.00
10000	0.07	0.00	0.00	10110	0.00	0.00	0.00
01000	0.02	0.00	0.00	11010	0.00	0.00	0.00
00100	0.00	0.00	0.00	11001	0.00	0.01	0.00
00010	0.01	0.01	0.04	10110	0.00	0.00	0.00
00001	0.00	0.07	0.00	10101	0.00	0.00	0.00
11000	0.01	0.00	0.00	10011	0.00	0.00	0.00
10100	0.00	0.00	0.00	01101	0.02	0.01	0.00
10010	0.00	0.00	0.00	01011	0.00	0.00	0.06
10001	0.02	0.00	0.00	00111	0.00	0.00	0.00
01100	0.02	0.00	0.00	11110	0.00	0.00	0.00
01010	0.00	0.00	0.04	11101	0.01	0.08	0.00
01001	0.01	0.06	0.00	11011	0.00	0.00	0.01
00110	0.00	0.00	0.00	10111	0.00	0.00	0.00
00101	0.00	0.00	0.00	01111	0.00	0.00	0.01
00011	0.00	0.00	0.01	11111	0.07	0.13	0.15

For five attributes, 32 ($2^5=2^5$) latent classes can be created. The highlighted characters in Table 7, which includes the rates of students' presence in the latent classes, indicated the most common five latent classes for each grade. When the table is examined, it is seen that the rate of students in the "00000" latent class, in other words, who had non-mastery for all attributes, was very high and close to each other for all three years. It was observed that approximately $\frac{3}{4}$ in 6th grade, in 7th and 8th grades $\frac{2}{3}$ of the students were in the "00000" latent class. At the 7th and 8th grades, the second largest latent class is "11111", with rates of 13% and 15%, respectively. This latent class consists of students who mastered all attributes. At the 6th grade level, those with all the attributes constituted the 6th largest group. Considering the ratios, it is seen that the number of students who mastered all the attributes was far behind the group sizes of those without any attributes. It is seen that the ratio was 0.00 in many latent classes. Many of these latent classes appeared to have no students due to the rounding process. However, it was observed that there were no students in some latent classes before the rounding process. It can be said that the students were not homogeneously distributed in the latent classes.

Rates of Change in Students' Mastery of Each Attribute

The latent class sizes contain a general result about the latent class in which students are included according to the measurement made in the relevant year. It can be examined in Table 8 which attributes of the students changed in the 7th grade compared to the 6th grade.

Table 8. Proportion of Students Whose Attribute Mastery Changes in 7th Grade According to 6th Grade based on Attribute

Change	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
Gainer	0.11	0.18	0.13	0.08	0.25
Loser	0.05	0.02	0.02	0.02	0.01
Total Change	0.17	0.20	0.14	0.10	0.26
Unchanging	0.83	0.80	0.86	0.90	0.73

When the attribute mastery status of the students as a result of the 7th grade measurements is compared with the results of the 6th grade from Table 8, it has been observed that the mastery status of approximately 4/5 of the students on the basis of the attribute did not change. The biggest change in the 7th grade was seen in attribute 5, in which 25% of the students gained the attribute and 1% lost. The smallest change was seen in attribute 4, where 8% of the students gained the attribute and 2% lost. When the changes are examined, it is seen that more students gained in all attributes. It can be examined in Table 9, which attributed mastery status of the students changed in the 8th grade when compared to the 7th grade.

Table 9. Proportion of Students Whose Attribute Mastery Changes in 8th Grade according to 7th Grade Based on Attribute

Change	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
Gainer	0.02	0.05	0.01	0.20	0.03
Loser	0.08	0.08	0.08	0.01	0.15
Total Change	0.10	0.13	0.09	0.21	0.18
Unchanging	0.90	0.87	0.91	0.79	0.82

When the attribute mastery status of the students as a result of the 8th grade measurement was compared with the results of the 7th grade Table 9, it was seen that the mastery status of more than 4/5 of the students did not change on the basis of attributes. The biggest change was observed in the 8th grade in which 19% of the students gained the attribute and 1% lost. The smallest change was seen in attribute 3, in which 1% of the students gained the attribute and 8% lost. When the changes are examined, it is seen that more students lost in all attributes except attribute 4.

The Rate of Change in the Number of Attribute Mastered by Students

In order to see the reflection of the changes given in Tables 7 and 8 to the number of attributes mastered, the changes on student basis should be monitored. The rates of students gaining or losing the attribute in the 7th grade according to their 6th grade results are given in Table 10.

Table 10. Attribute Mastery Change Rates in 7th Grade According to 6th Grade on Student Basis

	No Gain	Gain 1	Gain 2	Gain 3	Gain 4	Gain 5
No Lost	0.62	0.08	0.08	0.05	0.05	0.03
Lost 1	0.06	0.01	0	0	0	
Lost 2	0.01	0.01	0	0		
Lost 3	0	0	0			
Lost 4	0	0				
Lost 5	0					

Values in Table 10 showed that 62% of the students remained in the same latent class in the 7th grade when compared with the 6th grade. It is seen that 29% of the students gained attribute/attributes without losing the attributes they have, while 7% lost one or two attributes without gaining attributes. It can be said that the change in the attributes of students was more in the direction of gaining. The rate of attribute changing from 7th to 8th grades on student basis is given in Table 11.

Table 11. Attribute Mastery Change Rates in 8th Grade According to 7th Grade on Student Basis

	No Gain	Gain 1	Gain 2	Gain 3	Gain 4	Gain 5
No Lost	0.66	0.07	0.03	0.02	0	0.01
Lost 1	0.06	0.02	0.03	0.01	0	
Lost 2	0.05	0.01	0	0		
Lost 3	0.01	0	0			
Lost 4	0.01	0				
Lost 5	0					

Values in Table 11 indicated that 2/3 of the students remained in the same latent class in the 8th grade when compared with the 7th grade. It is seen that 13% of the students gained new attributes/attributes without losing their attributes, while 13% lost their attributes/attributes without gaining attributes. It is seen that the change was in the direction of gaining or losing attributes of the students was more limited and balanced in the 8th grade. Table 12 shows the correlations between the correct response rate and the attribute mastery probability of the students across years.

Arguments of Reliability and Validity Regarding the Analysis Results

Table 12. Correlations between Correct Response Rate and Attributes Mastery Probability by Years

	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
6-7	.63	.62	.65	.57	.57
7-8	.77	.77	.77	.61	.71
6. CRR-A	.78	.82	.82	.69	.84
7. CRR-A	.87	.89	.87	.77	.87
8. CRR-A	.87	.89	.86	.86	.89

Note: CRR-A Correlation between Correct Response Rate and Probability of Attribute Mastery

In the first two lines of Table 12, correlations between the attribute mastery probability of students calculated in consecutive years for each attribute were displayed, and in the last three lines correlations were found between the correct response rate and the attribute mastery probability of students in each year. The correlations found in the table were calculated with the Pearson coefficient, and all relationships were found to be significant at the $p < 0.01$ level. When the first line is examined, it is seen that the 6th and 7th grades attribute mastery probability was moderately correlated. The lowest correlation coefficient found in the table was found to be between 0.47 belonging to the attribute 4 mastery probability in these years. It is seen that the correlation coefficients regarding the attribute mastery probability of 7th and 8th grades are higher than 6th-7th. The correlation coefficient calculated for attribute 4 was again lower than the other attributes. Other correlations contained high levels of correlation meanings.

When the correlations between the correct responce rate and attribute mastery probability, which were carried out to examine the convergent validity of the analysis results, are examined, it is striking that the correlation coefficients were high. The correlation coefficients seen in the 6th grade were highly correlated. The values seen in the 7th grade were higher than the values seen in the 6th grade for all the attributes. In the 7th grade, it was observed that attribute 4 had lower than the other coefficients, again. When eighth-grade values are examined, higher correlation coefficients were observed the ones in previous years. It can be said that the relatively lower correlation coefficient observed for attribute 4 was not observed in the 8th grade values, and all correlation coefficients were close to each other. In Table 13, correct classification rates of students in terms of each attribute and latent classes in each class are given.

Table 13. Classification Accuracy

	Overall	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
6 th Grade	.73	.87	.92	.95	.94	.93
7 th Grade	.81	.96	.94	.97	.95	.91
8 th Grade	.88	.98	.96	.98	.94	.97

Table 13 shows that the lowest classification accuracy which can be taken as the reliability of findings, is in the 6th grade with 0.70; the classification accuracy increased to 0.81 in the 7th grade and 0.88 in the 8th grade. It was observed that the classification accuracy on the basis of attributes was higher than the values obtained for the whole latent class as expected. The average classification accuracy of the attributes was 0.91 for the 6th grade, 0.95 for the 7th grade, and 0.96 for the 8th grade. It is seen that the

correct classification rates on the basis of attributes increased over the years. The fact that the value of the 7th grade for attribute 4 was slightly lower than the value of the 8th grade is considered as an exception for this information. The high rates given provided important information about the accuracy of the classification resulting from the analysis.

DISCUSSION and CONCLUSION

In this study, it was aimed to monitor student achievement with cognitive diagnosis models (CDM). For this purpose, 2009 6th grade, 2010 7th grade, and 2011 8th grade HSPT mathematics test data were used. Analyses were carried out with the help of Q matrix developed by the experts. When the students' achievements are examined with their raw score in the exam, it is seen that there was an increase (0.34; 0.42) in the transition from the 6th to 7th grade, and a decrease in the transition from the 7th to the 8th grade (0.42; 0.33). On the basis of the HSPT results of a year ago, the opposite changes were observed in the study conducted by Yakar (2011). When acting according to the classical test theory and monitoring student achievements, different results can be obtained due to the effect of the item difficulty.

When it is aimed to obtain qualified and in-depth information about students' achievements, the use of cognitive diagnostic models can be a source of detailed information. The 6th, 7th, and 8th grade HSPT mathematics items were examined during the Q-matrix creation stage by experts. They decided that items require attributes called "Operational Applications" (Attribute 1), "Mathematization Applications" (Attribute 2), "Concept Calculations" (Attribute 3), "Concept Advanced Applications" (Attribute 4), and "Geometric Manipulation" (Attribute 5). Each exam may require specific attributes. Considering the purpose and results of HSPT, it can be said that it is not designed for cognitive diagnosis. In order to benefit from cognitive diagnosis at the highest level, there are many studies that pre-design questions to reveal the existence of qualifications that students should have (Akbay, Terzi, Kaplan, Karaarslan, 2017; de la Torre, van der Ark, & Rossi, 2017; Sorrel et al., 2016; Templin & Henson, 2006; Tjoe & de la Torre, 2014). However, exams with different purposes (Chen & Chen, 2016; Liu, Huggins-Manley & Bulut, 2018; von Davier, 2008) can be used later for cognitive diagnosis by retrofitting. It can be said that while developing the Q matrix in retrofitting studies, examining the AIC and BIC model data fit indexes, making decisions with consensus by experts, and examining the Q-matrix validity with software, are the factors that make the use of the test for cognitive diagnosis functional and meaningful in this study.

As a result of the analysis made using the Q matrices created, the attribute mastery probabilities were generally low, an increase in the transition from the 6th to 7th grade (0.16-0.26) and a partial decrease in the transition from the 7th to 8th grade (0.26-0.24) were seen. It can be said that the direction of the change (except for Attribute 4 in 8th grade) was similar to the change in the correct response rates of students over the years. It is thought that this situation may be related to the curriculum. Indeed, it is seen that the concepts at the 7th grade were designed as the application of the concepts addressed in the 6th grade, but there are concepts (irrational numbers, inequalities, etc.) that students encounter for the first time at the 8th grade. This opinion is supported by the study of Kablan, Baran, and Hazer (2013). In this study, it is stated that the behaviors targeted according to grade levels were at the comprehension level at the 6th and 8th grades and at the application level at the 7th grade.

In the trend of change in mastery probability of attribute 4, it was seen that the ratio increased slightly in the transition from the 6th to 7th grade, and there was a noteworthy increase (0.17-0.34) in contrast to the general change in the transition from the 7th to 8th grade. It can be thought that the items at the 7th grade were mostly the basic applications of the 6th grade concepts, and the 8th grade items were designed to cover previous learning.

CDMs basically classify individuals according to their attributes. Those with an attribute mastery probability of 0.5 and above were classified as attribute master and those below 0.5 were classified as non-master. When the attribute mastery probability obtained as a result of the analysis was transformed into the latent class, as expected, the largest latent class was realized as the "00000" group in which the students non-master any attributes. More than 60% of the students took part in this latent class in three years. This situation may mean that the students did not acquire the behaviors targeted in the curriculum

or that the exam does not have the quality to measure these behaviors. The next largest latent class was seen as the "11111" latent group in which the students had all the attributes. However, according to the 6th, 7th, and 8th grades, only 6%, 13%, and 15% of the students were in this latent class, respectively. Although there was no linear hierarchy among the qualifications, it was anticipated that the qualification in the higher group would correspond to a more advanced structure. Accordingly, due to our education system, it is an expected result that as the grade level increased, the probability of having qualifications and even higher-level qualifications would increase in students who encountered different concepts and question types. However, the fact that a student appeared to have qualifications at a grade level should not mean that the relevant student would maintain the same qualifications or have more of that qualification as the grade level increased. Qualifications were not directly subject or curriculum based. The nature of the questions, in which the learning outcomes required by the subject or curriculum were tried to be determined, indicated the mathematical qualifications of the student. For this reason, the properties of the questions selected to measure the learning outcomes in classifying students were crucial. Within the scope of the exams examined in this study, it is noteworthy that the questions for Qualification 1 in the 6th Grade, Qualification 3 in the 7th Grade, and Qualification 3 in the 8th Grade were predominant. According to this situation, one of the most expected learning outcomes from 6th grade students was to complete operational practices, whereas one of the most expected learning outcomes from 7th and 8th grade students was to perform the operations for the computational applications of the concepts. Considering the developmental characteristics of the students, although it was appropriate to expect the applications of the basic qualities of numbers from the 6th grade students, questions that support mathematical thinking beyond the application of operations were expected at the next grade levels. However, the current results did not reflect this expectation. Uğurel, Moralı, and Kesgin (2012) also support this result by stating that HSPT includes information transfer in 6th grade, routine operations in 7th grade, and questions at both knowledge transfer and routine operations level in 8th grade. It can be said that other latent group sizes differed according to grade levels. Şen and Arıcan (2015) conducted a cognitive diagnosis analysis based on TIMSS 2011 8th grade mathematics responses of Turkish students, and they found 13% mastery for all attributes and 1% non-mastery for all attributes. When many variables such as the number of attributes defined for the test, the measurement frequency of the attributes, the examination of the attributes in the same item, and the model used for analysis are partially or completely different, the results to be obtained can vary significantly. Although the findings obtained were specific to the study, the fact that the majority of the students had no attributes was one of the prominent results of the study.

On the basis of attributes, it has been observed not a big change was observed in students' attributes in the 7th grade when compared to the 6th grade. It has been observed that approximately 80% of the students in each attribute did not change. It was observed that the change in the students' attributes in the 8th grade was less than the previous year. It was observed that the status of the attributes mastery did not change between 80-90%. Another prominent result was that the change in attribute 4 in 8th grade was in the opposite direction with the changes in other attributes. Accordingly, while attributes 1, 2, 3, and 5 moved together in terms of the direction of change according to the years, attribute 4 changed in the opposite direction of the others.

When attribute mastery status changes over the years were examined on the basis of students, a little more than half of the students who did not lose or gain any attribute in the 7th grade were compared with the 6th grade. While 38% of the students gained/lost their attributes, it has been observed that most of these students gained new attributes/attributes. When the attributes they mastered in the 8th grade were compared with the 7th grade, it is observed that the change was more limited when compared to the previous year. While no change was observed in 2/3 of the students, it was observed that the number of students who gained and lost their attributes was close to each other. When the changes by years on the basis of students and attributes are examined together, it is concluded that the change seen in the 7th grade was more and more positive than the one seen in the 8th grade.

It has been observed that the attribute mastery probabilities had high correlation values for consecutive years. The high correlation value confirmed the conclusion that the stability of the measurements and the changes in attribute mastery probabilities were limited across years. The fact that the correlation

values between the 7th and 8th grades were higher than the correlation values between the 6th and 7th grades shows that the differentiation in the change by years was also reflected in the correlation values.

The correlation coefficients between the number of correct answers and the probability of having the qualifications of the students can be seen as the convergent validity coefficient (Li et al., 2020). These values were found to be high. Thus, it can be said that the obtained results have an argument of validity. It is noteworthy that these correlation values, which could be observed for three different years, generally increased over the years. The fact that the lowest values for all correlation values belonged to attribute 4 continuously can be considered as a reflection of the direction of the change in this attribute's being in the opposite change direction when compared to the other attributes. The primary factor that may cause this situation seems to be that the number of questions related to this qualification was higher in the 8th grade compared to other grade levels. Another factor is the natural consequence of seeing an advanced qualification such as advanced applications of the concept in 8th grade students. Another factor is the natural result of 8th grade students' having an advanced qualification such as advanced applications of the concept. Along with the advanced grade level, the vast knowledge of the students enables them to perform more complex operations on mathematical concepts beyond operational applications.

For CDM, the Q matrix is considered to be the basic element that reflects the design of the assessment tool and determines the quality of the feedback obtained from the assessment tool (Rupp & Templin, 2008). In order to increase the robustness of the CDM results, experts and statistical validation opportunities were used in creating a Q matrix. The accuracy of the classification rates revealed at the end of the analysis was 70% for the 6th grade, 81% for the 7th grade, and 88% for the 8th grade. The accuracy rates on the basis of attributes were found between 87% and 97%. It can be said that the analysis produces more accurate results over the years. In the studies conducted (de la Torre, Yong, & Deng, 2010; Madison & Bradshaw, 2015), no threshold value for classification accuracy is specified. However, it is seen that the classification accuracy revealed in the study has a higher level than similar studies based on real data (Cui, Gierl, & Chang, 2012; Li et al., 2020; Ma, Iaconangelo, & de la Torre, 2016). The high rates obtained reveal the accuracy of the analysis results and indicate that the comments made on the results can be trusted.

It is among the limitations of this study that HSPT did not have a diagnostic purpose and therefore did not have a predetermined Q matrix. Analysis of the data with a purpose or method other than its original purpose or analysis method is called retrofitting, and potential problems such as model-data fit fatigue may be encountered. Although it is desirable to prepare items based on the Q matrix, there are many studies performed through retrofitting (Chen & Chen, 2016; Liu, Huggins-Manley & Bulut, 2018; von Davier, 2008). The fact that the research data belongs to the previous years can be seen as a limitation. However, the same person in the succession of tests is limited, and HSPT was the only repeated measure high-stake exam for Turkey. There are suggested models in the analysis of longitudinal data with CDM (Huang, 2017; Kaya & Leite, 2017; Zhan et al., 2019). However, it was not possible to use it in this study since all of the suggested models are based on the common item.

It is one of the main advantages of CDMs that they provide detailed information about individuals. The fact that this benefit is also for monitoring student development makes CDMs more functional in evaluation. With the use of CDMs in large-scale exams, the knowledge that students get from the exams will not be limited to the correct numbers they make. By determining what level of deficiencies in which skill they have, the first step will be taken towards making up these deficiencies of students. Other stakeholders in education, such as the school, decision-makers, and parents, will also have the option to act on these deficiencies. When this feature is transferred to the exams held in series, student progress can be examined over the years over common attributes, as shown in the study. In this context, CDMs can be used for monitoring purposes in schools. Based on research results, suggestions for researchers are as follows;

- The research had a design in which the Q matrix was subsequently determined. In future studies, if the Q matrix is determined in advance and the items are created based on this, the classification accuracies can be examined.

- Analysis was done repeatedly due to the lack of a suitable growth model. In particular, existing models can be developed to analyze data used in research at once.
- If such a model is developed, the research data can be analyzed again, and the attribute mastery probabilities can be examined.
- The learning outcomes aimed within the scope of the curriculum are subject-curricular-based and remain only within their own context, and it cannot be examined to what extent the students acquire the skills required by these learning outcomes. For this reason, it is not possible to observe the qualifications properly at all levels. It will be more meaningful to determine the skills expected from students in such leveling exams in advance and to create questions in the context of these skills in order to determine the qualifications of the student who will proceed to the next level.

REFERENCES

- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, 19, 716–723. doi:10.1109/TAC.1974.1100705
- Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K. G. (2017). Expert-based attribute identification and validation on fraction subtraction: A cognitively diagnostic assessment application. *Journal on Mathematics Education*, 9(1), 103-120.
- Chen, H., ve Chen, J., (2016). Retrofitting Non-cognitive-diagnostic Reading Assessment Under the Generalized DINA Model Framework, *Language Assessment Quarterly*, 13(3), 218-230, DOI: 10.1080/15434303.2016.1210610
- Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70 (6), 902-913.
- Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, 49(1), 19-38.
- de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement*, 33(3), 163–183.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179-199.
- de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, 47(2), 227-249.
- de la Torre, J., & Minchen N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, 20(2), 89-97.
- de la Torre, J., van der Ark, L. A., & Rossi, G. (2017). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 1-16.
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 979-1029). Amsterdam: Elsevier
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380-396
- Hartz, S. M. (2002). *A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*, Unpublished PhD dissertation, University of Illinois at Urbana-Champaign, ABD.
- Huang, H. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement*, 54: 440-480. doi: 10.1111/jedm.12156
- Iaconangelo, C. (2017). *Uses of classification error probabilities in the three-step approach to estimating cognitive diagnosis models*. (Unpublished doctoral dissertation). New Brunswick, NJ: Rutgers University.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258-272.
- Kablan, Z., Baran, T. & Hazer, Ö. (2013). İlköğretim matematik 6-8 öğretim programında hedeflenen davranışların bilişsel süreçler açısından incelenmesi. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, 14 (1), 347- 366.
- Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement*, 77(3), 369-388.
- Kutlu, Ö., Doğan, C., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme*. Pegem, Ankara

- Li, L., Zhou, X., Huang, J., Tu, D., Gao, X., Yang, Z., & Li, M. (2020). Assessing kindergarteners' mathematics problem solving: The development of a cognitive diagnostic test. *Studies in Educational Evaluation*, 66, 100879.
- Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from irt-based assessment forms. *Educational and Psychological Measurement*, 78(3), 357–383. <https://doi.org/10.1177/0013164416685599>
- Leighton, J. P. ve Gierl M. J. (2007). Why cognitive diagnostic assessment, Leighton, J. P. Gierl M. J. (Eds). *Cognitive Diagnostic Assessment for Education*. Cambridge University Press, New York, USA.
- Lohman, D. F. (1999). Minding our p's and q's: On finding relationships between learning and intelligence. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *The future of learning and individual differences: Process, traits, and content* (55f72). Washington, DC: American Psychological Association.
- Ma, W. & de la Torre, J. (2019). GDINA: The generalized DINA model framework. R package version 2.3. Retrieved from <https://CRAN.R-project.org/package=GDINA>
- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40(3), 200-217.
- Madison, M. J., & Bradshaw, L. P. (2015). The Effects of Q-Matrix Design on Classification Accuracy in the Log-Linear Cognitive Diagnosis Model. *Educational and Psychological Measurement*, 75(3), 491–511. <https://doi.org/10.1177/0013164414539162>
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187-212.
- Ministry of National Education (MoNE) (2018). *Matematik dersi öğretim programı (İlkokul ve Ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. Sınıflar)*. Ankara: Talim ve Terbiye Kurulu Başkanlığı.
- Ministry of National Education (MoNE), (2008), *64 Soruda Ortaöğretime Geçiş Sistemi ve Seviye Belirleme Sınavı Örnek Sorular*. Ankara: MEB Yayınları
- Ministry of National Education (MoNE), (2010), *Seviye belirleme sınavının değerlendirilmesi*. Turkish Ministry of National Education, Retrieved from https://www.meb.gov.tr/earged/earged/sbs_deger.pdf
- Ministry of National Education (MoNE), (2017), *Akademik becerilerin izlenmesi ve değerlendirilmesi: 8. Sınıflar raporu*. Turkish Ministry of National Education, Retrieved from http://edirne.meb.gov.tr/meb_iys_dosyalar/2018_06/08104327_ABYDE_Turkiye.pdf.
- Organisation for Economic Co-operation and Development (OECD), (2019), *PISA 2018: Insights and Interpretations.*, Retrieved from <https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF..>
- Rupp, A. A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68(1), 78–96. <https://doi.org/10.1177/0013164407301545>
- Schwarzer, G. (1976). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464. doi:10.1214/aos/1176344136
- Sezen Yüksel, N., Sağlam Kaya, Y., Urhan, S. & Şefik, Ö. (2019). *Matematik Eğitiminde Modelleme Etkinlikleri (Ed: Şenol Dost)*. Ankara: Pegem Akademi
- Smith, G.H., Wood, L.N., Coupland, M., Stephenson, B., Crawford, K. & Ball, G. (1996). Constructing mathematical examinations to assess a range of knowledge and skills. *Int. J. Math. Educ. Sci. Technol.*, 27(1), 65-77.
- Sorrel, M., Olea, J., Abad, F., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*. 19(3), 506-532, doi: 10.1177/1094428116630065
- Şen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, 6(2), 238-253.
- Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901-926.
- Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26, 237-255.
- Uğurel, I., Morali, H.S. & Kesgin, Ş. (2012). OKS, SBS ve TIMSS matematik sorularının 'math taksonomi' çerçevesinde karşılaştırmalı analizi. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi*, 11 (2), 423- 444.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287–307. doi:10.1348/000711007X193957

- Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: a higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, 43(1), 57-87.
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52, 457-476
- Yakar, L. (2011). *İlköğretim ikinci kademe öğrencilerinin SBS puanları ve akademik başarı puanları değişimlerinin izlenmesi ve SBS puanlarının kestirilmesi*. Unpublished Master dissertation. Abant İzzet Baysal Üniversitesi /Eğitim Bilimleri Enstitüsü, Bolu, Turkey.
- Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, 44(3), 251-281.