Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Bahar* 2017 *Spring* 2017 *Cilt: 8- Sayı: 1* *Volume: 8- Issue: 1*

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)

Editör
Prof. Dr. Selahattin GELBAL

Yardımcı Editör
Yrd. Doç. Dr. Kübra ATALAY KABASAKAL
Dr. Sakine GÖÇER ŞAHİN

Genel Sekreter
Doç. Dr. Tülin ACAR

Yayın Kurulu
Prof. Dr. Terry A. ACKERMAN
Prof. Dr. Cindy M. WALKER
Doç. Dr. Cem Oktay GÜZELLER
Doç. Dr. Neşe GÜLER
Doç. Dr. Hakan Yavuz ATAR
Doç. Dr. Oğuz Tahsin BAŞOKÇU
Yrd.Doç.Dr. Hamide Deniz GÜLLEROĞLU
Yrd. Doç. Dr. Derya ÇOBANOĞLU AKTAN
Yrd. Doç. Dr. Okan BULUT
Yrd. Doç. Dr. N. Bilge BAŞUSTA
Dr. Nagihan BOZTUNÇ ÖZTÜRK

Dil Editörü
Doç. Dr. Burcu ATAR
Yrd. Doç. Dr. Derya ÇOBANOĞLU AKTAN

Sekreterya
Arş. Gör. İbrahim UYSAL
Arş. Gör. Seçil UĞURLU
Arş. Gör. Nermin KIBRISLIOĞLU UYSAL

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

İletişim
e-posta: epod@epod-online.org
Web: http://epod-online.org

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

Editor
Prof. Dr. Selahattin GELBAL

Assistant Editor
Assist. Prof. Dr. Kübra ATALAY KABASAKAL
Dr. Sakine GÖÇER ŞAHİN

Secratary
Doç. Dr. Tülin ACAR

Editorial Board
Prof. Dr. Terry A. ACKERMAN
Prof. Dr. Cindy M. WALKER
Assoc. Prof. Dr. Cem Oktay GÜZELLER
Assoc. Prof. Dr. Neşe GÜLER
Assoc. Prof. Dr. Hakan Yavuz ATAR
Assoc. Prof. Dr. Oğuz Tahsin BAŞOKÇU
Assist. Prof. Dr. Hamide Deniz GÜLLEROĞLU
Assist. Prof. Dr. Derya ÇOBANOĞLU AKTAN
Assist. Prof. Dr. Okan BULUT
Assist. Prof. Dr. N. Bilge BAŞUSTA
Dr. Nagihan BOZTUNÇ ÖZTÜRK

Language Reviewer
Assoc. Prof. Dr. Burcu ATAR
Assist. Prof. Dr. Derya ÇOBANOĞLU AKTAN

Secretarait
Res. Assist. İbrahim UYSAL
Res. Assist. Seçil UĞURLU
Res. Assist. Nermin KIBRISLIOĞLU UYSAL

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

Contact
e-mail: epod@epod-online.org
Web: http://epod-online.o

Dizinleme / Abstracting & Indexing
DOAJ (Directory of Open Access Journals), TÜBİTAK Ulakbim Sosyal ve Beşeri Bilimler Veri Tabanı, Tei (Türk Eğitim İndeksi)

Hakem Kurulu / Referee Board

Adnan KAN (Gazi Üni.)
Ahmet TURAN (Pearson)
Ali BAYKAL (Bahçeşehir Üni.)
Adnan ERKUŞ (Emekli Öğrtim Üyesi)
Arif ÖZER (Hacettepe Üni.)
Ayfer SAYIN (Gazi Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Bayram BIÇAK (Akdeniz Üni.)
Bayram ÇETİN (Gazi Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Bilge GÖK (Hacettepe Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Beyza AKSU DÜNYA (Illinois Uni.)
Cem Oktay GÜZELLER (Hacettepe Üni.)
Cindy M. WALKER (Duquesne University)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Dilara BAKAN KALAYCIOĞLU (ÖSYM)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu GÜNGÖR (İzmir Üni.)
Elif Bengi ÜNSAL ÖZBERK (Adalet Bakanlığı)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Eren Halil Özberk (Hacettepe Üni.)
Ergül DEMİR (Ankara Üni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Hacettepe Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Evrim ÇETİNKAYA YILDIZ (Erciyes Üni.)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Göksu GÖZEN (Mimar Sinan Güzel Sanatlar Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Sakarya Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hüseyin SELVİ (Mersin Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kaan Zülfikar DENİZ (Ankara Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Hacettepe. Üni.)
Mehmet KAPLAN (MEB)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Mustafa ASİL (University of Otago)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Neşe GÜLER (Sakarya Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Ankara Üni.)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Hacettepe Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (Hacettepe Üvi.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Sema SULAK (Bartın Üni.)
Serdar ÇAĞLAK (Osmangazi Üniveristesi)
Seval KIZILDAĞ (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of North Carolina)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)

Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2017, Sayı: 8-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2017, Issue: 8-1)

İÇİNDEKİLER / CONTENTS

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575

Sayın okurlar,

Bu sayı itibariyle, yılda iki kez olmak üzere ilk sayısını 2010 yılında yayınladığımız Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi'ni yılda dört kez yayınlamanın gururunu ve heyecanını yaşamaktayız. Bu başarıyı hiç şüphesiz derginin kuruluşundan bu yana görev alan editörlerin, bilim kurulunun, hakemlerin katkısına ve siz okurlarımızın ilgisine borçluyuz.

Dergimize gönderilen aday makalelerin sayısının artması derginin iş hacmini de arttırmaktadır. Bu sebeple hem editör ekibinde hem sekreteryada hem de bilim kurulunda görev alan arkadaşlarımızın sayısını arttırma ihtiyacı duyduk. Bu ilk sayı, yeni ve güçlü bir ekibin birlikte çalışmasının ürünü olup, bundan sonraki her bir sayının bir öncekinden daha iyi olacağından eminiz.

Dergimiz, 2014 yılı itibariyle ULAKBİM ve 2016 yılı itibariyle DOAJ indeksinde taranmaktadır. Ayrıca 2017 yılının başında dergimizin SCOPUS indeksinde taranması için başvurumuzu gerçekleştirmiş bulunuyoruz. Bundan sonraki hedefimiz, dergimizi uluslararası indeksli bir dergi haline getirmektedir. Bu amaçla gerekli koşulları yerine getirmek için çalışmalarımız devam etmektedir.

Bu sayıda, on değerli makalenin yanı sıra Prof. Dr. Stephen G. SIRECI ile geçerlik üzerine yapılan söyleşinin tam metni yer almaktadır. Dergimize gösterdiğiniz ilgiye ve ölçme ve değerlendirme alanına kattığınız değere teşekkür eder, keyifli okumalar dileriz.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi Editör Ekibi

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);1-14

The Big Fish-Little Pond Effect on Affective Factors Based on PISA 2012 Mathematics Achievement

PISA 2012 Matematik Başarısına Dayalı Duyuşsal Faktörlerde Büyük Balık-Küçük Göl Etkisi

Dilara BAKAN KALAYCIOĞLU*

Abstract
In this study, the 2012 PISA Turkey student questionnaire data is considered to determine the big fish-little pond effect. The mathematics self-efficacy, self-concept and anxiety affective factors are examined to explain the relation of each of them with the school type, gender, socioeconomic status, student's mathematics achievement and school's mathematics achievement covariates. A total number of 771 students from 88 high schools are in the sample. Factor analyses' results support the construct validity of the Student Questionnaire's mathematics self-efficacy, anxiety and self-concept items. Data set is analyzed with Multiple Indicator Multiple Cause Model and the patterns of association with covariates and affective factors were tested simultaneously. According to the results, Anatolian high school students have a higher mathematics self-efficacy and lower mathematics anxiety than do the general high school students. However, when the student mathematics achievement and school mathematics achievement variables were inserted to the model, school type was not associated with mathematics self-efficacy. Moreover, Anatolian high school student's mathematics anxiety was higher than that of the general high school students. Student's mathematics achievement was the most significant predictor of the mathematics self-efficacy, anxiety and self-concept factors. Finally, school's mathematics achievement was a significant predictor of only mathematics self-concept. The identification of increase in school's mathematics achievement yields a decrease in the student's mathematics self-concept may be considered as the most important result of this study. These results provide evidence about the Anatolian high schools' students experience big fish-little pond effect.

Keywords: Big fish-little pond effect, mathematics self-efficacy, anxiety, self-concept

Öz
Bu çalışmada, PISA 2012 Türkiye öğrenci anketi kullanılarak büyük balık-küçük göl etkisi belirlenmeye çalışılmıştır. Matematik dersine yönelik öz-yeterlik, benlik ve kaygı duyuşsal faktörlerinin okul türü, cinsiyet, sosyoekonomik statü, öğrenci matematik başarısı ve okul matematik başarısı değişkenleri ile arasındaki ilişkiler incelenmiştir. Örneklemde 88 liseden 771 öğrenci bulunmaktadır. Açımlayıcı ve doğrulayıcı faktör analizi sonuçları öğrenci anketindeki matematik öz-yeterlik, benlik ve kaygı maddelerinin yapı geçerliğini desteklemektedir. Analizler, Çoklu Gösterge ve Çoklu Neden Modeli ile gerçekleştirilmiş ve model yardımıyla neden değişkenleriyle, duyuşsal faktörler arasındaki ilişki örüntüleri aynı anda kestirilmiştir. Sonuçlara göre, Anadolu lisesi öğrencileri genel lise öğrencilerinden daha yüksek matematik öz-yeterlik algısına ve daha düşük matematik kaygısına sahiptir. Ancak, modele öğrenci matematik başarısı ve okul matematik başarısı değişkenleri de eklendiğinde, okul türü matematik öz-yeterliği ile ilişkili değildir. Ayrıca, Anadolu lisesi öğrencilerinin matematik kaygısının genel lise öğrencilerinden daha yüksek olduğu belirlenmiştir. Matematik öz-yeterlik, benlik ve kaygı değişkenlerinin en güçlü tahmin edicisi, öğrenci matematik başarısıdır. Son olarak, okul matematik başarısı sadece matematik benlik algısının istatistiksel olarak anlamlı bir tahmin edicisidir. Bu çalışmanın en önemli bulgusu, okul matematik başarı ortalaması arttıkça öğrencilerin matematik benlik algılarının düşme eğiliminde olduğunun belirlenmesidir. Bu sonuçlar, büyük balık-küçük göl etkisinin Anadolu lisesi öğrencileri tarafından yaşandığına dair kanıt sunmaktadır.

Anahtar Kelimeler: Büyük balık-küçük göl etkisi, matematik öz-yeterlik, kaygı, benlik

* Dr., Ölçme Seçme ve Yerleştirme Merkezi, Soru Hazırlama ve Geliştirme Daire Başkanlığı, Ankara-Türkiye, e-posta: dilara.bakan@osym.gov.tr

DOI: 10.21031/epod.297686
Received: 04.11.2016
Accepted: 31.01.2017

INTRODUCTION

Relationship between affective factors and academic achievement is one of the most widely studied subjects in the literature. There is a reciprocal relationship between affective factors and academic achievement (Marsh, Hau, Artelt, Baumert & Peschar, 2006; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005). According to the social cognitive theory, self-efficacy is defined as the perception of one's own ability to complete tasks and reach goals (Bandura, 1993), which is one of these affective factors. Self-concept is another affective factor that is described as the students' beliefs about themselves that include own knowledge, value, abilities and aims (Hattie, 1992). Thus, mathematics self-concept is defined as the student's beliefs in their perception of being successful at mathematics and their confidence as mathematics learners (Reyes, 1984; Wilkins, 2004). Contrary to self-concept, mathematics anxiety is the feeling of helplessness and stressful when dealing with mathematics (Ashcraft, 2002; Hembree, 1990). Different studies show that self-efficacy and self-concept are positively (Alcı, Erden & Baykal, 2010; Chiu & Xihua, 2008; İş Güzel & Berberoğlu, 2010; Hammouri, 2010; Özel, Çağlak & Erdoğan, 2012; Wilkins, Zembylas & Travers, 2002), and anxiety is negatively related to students' academic performance (Cassady & Johnson, 2002; Ho et al., 2000; Ma, 1999; McCarthy & Goffin, 2005; Seipp, 1991).

Gender and socioeconomic status are also influential variables on students' affective factors. It was shown that generally male students' mathematics self-efficacy (Bong, 1998; Özgen & Bindak; 2011; Pajares, 1996; Taşdemir, 2012) and mathematics self-concept (Pajares & Miller; 1994) are higher than that of female students'. When the relationship between mathematics anxiety and gender is discussed, some research studies show that female students have more anxiety than do the male students (Baloglu & Kocak, 2006; Hembree, 1990; Jain & Dowson, 2009; Wigfield & Meece, 1988) whereas other studies indicate that there is no significant difference between the gender and the mathematics anxiety (Cooper & Robinson, 1989; Pajares & Graham, 1999; Kurbanoğlu & Takunyacı, 2012). Some authors show that students with higher socioeconomic status have higher mathematics self-efficacy and self-concept perception than do the students with lower economic status (Marsh et al., 2006) and these students also have a lower mathematics anxiety than do the students with lower socioeconomic status (Geist, 2010).

In addition to the mentioned variables above, the social comparison environment can also influence the affective factors. According to the social comparison theory, people evaluate their opinions and abilities by comparing themselves with other people around their environment (Festinger, 1954). Therefore, schools are one of these social comparison environments that influence our self-evaluation. The effect of social comparison, especially on the self-concept, is well-known (Wood, 1989). In the literature, there are evidences about the negative impacts of academically successful student's getting education with the similar level of students on the self-concept perception (Marsh, Trautwein, Lüdtke, Baumert & Köller, 2007). The students who do grouped according to their academic capabilities have lower self-concept perception than the similar students who are at a combined talent level school (Nagengast & Marsh, 2011), which is generally called as big fish-little pond effect. Big fish-little pond effect is especially encountered at the education systems where the students are grouped according to their academic success (Ho, 2009). Big fish-little pond effect was also shown in the studies performed by the data of Germany (Marsh, Köller & Baumert, 2001), Hong Kong (Marsh, Kong & Hau, 2000), Israel (Zeidner & Schleyer, 1999), and Australia (Marsh, Chessor, Craven & Roche, 1995). Altun and Yazıcı (2013) also explain the higher self-concept perception of vocational, general and Anatolian high school students than that of the science high school students with the big fish-little pond effect.

On the other hand, in general, the studies about self-efficacy and anxiety show that the students with high ability levels have higher mathematics self-efficacy than do the students with the low ability levels (Pajares & Graham, 1999). In Turkey, Anatolian high school student's mathematics self-efficacy is higher than that of the general high school students (Kurbanoğlu & Takunyacı, 2012; Özgen & Bindak, 2011) whereas their anxiety levels are lower than that of the general high school students (Kurbanoğlu & Takunyacı, 2012).

In Turkey, there was a centralized multiple-choice examination for the student selection to the high schools except the general high schools in 2012. While the students who are admitted to the Anatolian high schools compare themselves with the similar ability level students, the general high school students usually have the chance to compare themselves with a high range of different ability level students. Thus, the social comparison environment changes according to the school types. Since the students are grouped according to their academic achievement in Turkey, the influence of big fish-little pond effect is encountered at the school level. Thus, it is believed that the student's awareness affects their perception of mathematics self-efficacy, anxiety and self-concept.

There are many studies about the relation of the socioeconomic status, school type, gender and student's mathematics achievement with the mathematics self-efficacy, anxiety and self-concept. To the best knowledge of author, there is not any study that considers the school's mathematics achievement with these mentioned variables and the effects of big fish-little pond on the student's mathematics self-efficacy, anxiety and self-concept for Turkish data.

In this study, the 2012 data of Programme for International Student Assessment (PISA), for Turkey is considered to determine the big fish-little pond effect for Anatolian high school students and general high school students. Thus, the mathematics self-efficacy, self-concept and anxiety factors are examined to explain the relation of each of them with the gender, socioeconomic status, student mathematics achievement, and school mathematics achievement. The aim of this study is to find answers to the following questions.

1. Are there any statistical difference between the Anatolian and general high school students' mathematics self-efficacy, anxiety and self-concept in terms of school type, gender and socioeconomic status?

2. Are there any statistical difference between the Anatolian and general high school students' mathematics self-efficacy, anxiety and self-concept in terms of school type, gender and socioeconomic status, mathematics achievement and school's mathematics achievement?

METHOD

In this study, PISA 2012 Turkish data set is analyzed with MIMIC (Multiple Indicator Multiple Cause) model which is a special case of structural equation modeling. MIMIC model consists of two parts; a measurement model which defines the relations between a latent variable and its indicators and a structural model which specifies causal relationships among latent variables and explains the causes (covariates) (Jöreskog & Sörbom, 1996). In measurement model, latent variables are defined by observable indicator variables. Latent variables of this study are mathematics self-efficacy, anxiety, self-concept; observable indicator variables are PISA student questionnaire items; covariates are school type, gender, socioeconomic status, student mathematics achievement and school mathematics achievement. MIMIC modeling was used in order to assess the effect of covariates on the mentioned three affective factors. MIMIC model provides simultaneous detection of association between the covariates and latent variables and it also indicates the unique effects of each covariate after controlling for the effects of the other covariates.

Sample

The data analyzed in this research were obtained from the PISA 2012 assessment. The target population of PISA 2012 was fifteen years old students. 4848 students from Turkey are participated in PISA 2012. PISA employed two-stage sampling procedure to ensure that a representative sample of target population for each country (OECD, 2013). Different students take different combinations of items and student questionnaire. We only consider students who completed the mathematics self-efficacy, anxiety and self-concept scales in Turkish data set. A total number of 771 students (423 females, 348 males) from 88 high schools (36 Anatolian and 52 general high schools) are in the

sample. Among these students, 333 of them are (42%) Anatolian high school and 438 (58%) of them are general high school students. The missing values were replaced by mean values of the variables.

Measures

The data set were acquired from the official website of OECD. The PISA 2012 study consisted of achievement tests along with a student questionnaire, a teacher questionnaire and a school principal questionnaire. In th's present study, items from the student questionnaire associated with school type, gender, socioeconomic status, self-efficacy, anxiety and self-concept are used. These items are explicitly described below.

Anatolian and general high schools are the two types of high schools considered in this study. Anatolian high schools are public schools that admit their students according to high stake nation-wide standardized examination. General high schools are also public schools but that admit students without examination. The PISA code of the Anatolian high schools is 07920002, which is coded as 1 whereas the PISA code of the general high schools is 07920003, which is coded as 0. Male and female students are coded as 1 and 0, respectively.

Socioeconomic status *(SES)* in PISA is assessed by educational, social and cultural status (ESCS) index. ESCS index is a composite index of three separate indexes: highest parental education, highest parental occupation, and cultural economic resources. The index maintains a mean of 0 and a standard deviation of 1 for students from OECD countries (OECD, 2013).

PISA aims to measure students' ability to handle certain mathematics processes originating from a real world context. Mathematics achievement test contains four content categories: quantity, uncertainty and data, change and relationship, space and shape. Some questions are multiple choice and others require students to construct their response. Since the PISA assessment used incomplete assessment design, students are required to answer a subset of the item battery; PISA estimates students' test scores as plausible values with each student having five plausible values for mathematics performance. Plausible values represent the range of abilities that a student might reasonably have based on student responses to the subset of items they receive, as well as on other relevant and available background information (Wu & Adams, 2002). PISA calculated mathematics achievement scores of students with a mean of 500 and standard deviation of 100. Students' mathematics achievement scores (five plausible values) are converted to z scores. A separate data analyses with mathematics achievements was run for each plausible value (PV1MATH-PV5MATH) and in order to provide unbiased estimates, the results are averaged according to the OECD (2009) protocols. School mathematics achievement scores are calculated with the school's students' mathematics achievement mean scores and then these scores are converted to z scores.

Mathematics self-efficacy items asked students to report on their confidence in doing a range of mathematical tasks with an example is using a 4-point agreement scale. Students' responded eight items, whether they feel "very confident", "confident", "not very confident" and "not at all confident". The highest response code was indicative of a positive rating of self-efficacy.

Mathematics anxiety items include five statements and asked students to report what extent they are agreed to given statements is using a 4-point agreement scale. Responses are ranged between "strongly agree", "agree", "disagree" and "strongly disagree". Similar to previous scale, the highest response code was indicative of a positive rating of mathematics anxiety.

Mathematics self-concept items include five statements and asked students to report what extent they are agreed to given statements is using a 4-point agreement scale (OECD, 2013). Responses are ranged between "strongly agree", "agree", "disagree" and "strongly disagree".

Analysis of Data

In order to examine factor structure, the selected items were subjected to Exploratory Factor Analysis (EFA). Based on factor analysis results, items which loaded different factors for the different school type is excluded from analysis. Thus, four of the original 18 items were not considered for further EFA. Cronbach's alpha (α) correlation coefficient value was also calculated for each factor using reliability analysis. Prior to the insertion of the covariates, Confirmatory Factor Analysis (CFA) were performed to evaluate construct validity of measurement model. Then, MIMIC model applied to compare path coefficients of Anatolian and general high schools.

The indices used to test goodness of model fit are RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), NFI (Normed Fit Index) and GFI (Goodness of Fit Index). Lower than .05 RMSEA value indicates excellent fit; the RMSEA value around .08 indicates adequate fit. Larger than .95 of CFI, NFI, GFI values reflect a good fit (Hu & Bentler, 1998; Schumacker & Lomax, 2010). χ^2 test statistics is directly affected by sample size and for large samples trivial differences may become significant. χ^2 test statistics are also provided but due to its sensitivity to sample size to assess model fit, RMSEA, CFI, NFI and GFI fit indices are interpreted. (Browne & Cudeck, 1993; Wu, Li, & Zumbo, 2007).

All analyses reported in this study are carried out using LISREL 8.80 for Windows (Jöreskog & Sörbom, 2006) with SIMPLIS command language. Since the mathematics self-efficacy, anxiety and self-concept items are categorical, WLS (Weighted Least Square) estimation method is used in confirmatory factor analysis and RML (Robust Maximum Likelihood) estimation method is used in MIMIC models analyses.

RESULTS

Means and standard deviations of variables for different school types are presented in Table 1.

Table 1. Means and Standard Deviations of Variables for Different School Types

School type	SES		Student mathematics achievement		School mathematics achievement		Mathematics self-efficacy		Mathematics anxiety		Mathematics self-concept	
	$\bar{X}$	*sd*	$\bar{X}$	*sd*	$\bar{X}$	*sd*	$\bar{X}$	*sd*	$\bar{X}$	*sd*	$\bar{X}$	*sd*
General	-1.68	.48	414.80	63.46	412.08	55.04	17.68	3.37	10.27	2.93	9.63	3.18
Anatolian	-.82	.64	530.11	70.04	521.55	48.54	19.29	3.08	9.59	2.87	9.88	3.02

As given in Table 1, Anatolian high school students' means are higher than that of the general high school students for all variables except mathematics anxiety. The correlations between variables are presented in Table 2.

Table 2. Correlations between Variables

Variables	1	2	3	4	5	6	7	8
1. School type	1							
2. Gender	-.03	1						
3. SES	.38*	.01	1					
4. Student mathematics achievement	.66*	.07	.41*	1				
5. School mathematics achievement	.78*	-.03	.47*	.81*	1			
6. Mathematics self-efficacy	.24*	.16*	.23*	.40*	.32*	1		
7. Mathematics anxiety	-.11*	.02	-.10*	-.28*	-.18*	-.23*	1	
8. Mathematics self-concept	.04	.08	.10*	.21*	.08	.42*	-.52*	1

Note. N=771. *p<.01.

Correlation between the variables showed wide range of values indicating varying levels of strength of associations. Patterns of correlation were generally within the theoretical expectations. As can be seen in Table 2, the highest correlation value of .81 is between mathematics achievement and school mathematics achievement. The highest negative correlation value of -.52 is between mathematics anxiety and mathematics self-concept.

In the first step, the principal axis factoring with items from the student questionnaire associated with Varimax rotation method was used to examine the underlying structure of the data. The factor loadings from EFA presented in Table 3.

Table 3. Factor Loadings of the Exploratory Factor Analysis with Varimax Rotation

Items	Factor		
	1	2	3
Mathematics self-concept			
St42Q07. I have always believed that mathematics is one of my best subjects.	**.76**	.20	-.25
St42Q06. I learn mathematics quickly.	**.70**	.25	-.29
St42Q04. I get good marks in mathematics.	**.68**	.17	-.25
St42Q09. In my mathematic class. I understand even the most difficult work.	**.67**	.26	-.18
Mathematics self-efficacy			
St37Q02. Calculating how much cheaper a TV would be after a 30% discount	.07	**.70**	-.11
St37Q01. Using a train timetable to work out how long it would take to get from one place to another	.10	**.64**	-.05
St37Q03. Calculating how many square meters of tiles you need to cover a floor	.24	**.63**	-.09
St37Q08. Calculating petrol consumption rate of a car	.18	**.54**	-.02
St37Q04. Understanding graphs presented in newspapers	.05	**.47**	-.05
St37Q06. Finding the actual distance between two places on a map with a 1:10.000 scale	.19	**.40**	-.08
Mathematics anxiety			
St37Q03. I get very tense when I have to do mathematics homework	-.30	-.10	**.75**
St37Q05. I get very nervous doing mathematics problems	-.09	-.09	**.71**
St37Q08. I feel helpless when doing a mathematics problem	-.19	-.09	**.68**
St37Q01. I often worry that it will be difficult for me in mathematics classes	-.37	-.06	**.61**

Note. ST42Q07 refers to question 42, stem 7 in students' questionnaire.

As given in Table 3, EFA derived three factors, mathematics self-concept, self-efficacy, and anxiety; their corresponding eigen values are 4.88, 2.08 and 1.25, respectively. These factors were named according to the common characteristics of the items loaded on the same factors, and also based on PISA 2012 framework (OECD, 2013). These three factors explained 58.62% of the total variances. Total variances explained by mathematics self-concept, self-efficacy and anxiety factors are 34.86%, 14.85% and 8.91%, respectively. Four items loaded on the mathematics self-concept factor (α=.86), six items on mathematics self-efficacy factor (α=.74), and four items on mathematics anxiety factor (α=.82). Factor loadings of items take values between .40 and .76, which is satisfactory.

In the second step, CFA were carried out to confirm the EFA results. The CFA model for Anatolian high schools yielded a $\chi^2(74)= 150.10$, $p<.0001$, RMSEA=.037 [.028:.045], CFI=.97, NFI=.95, GFI= .99. Overall model fit seems adequate based on values of selected fit indexes. These results indicated that, the three factors model that emerged in the EFA was confirmed by the CFA.

In the third step, the structural part of the model was inserted to measurement model and the MIMIC model was estimated. The influence of school type, gender, and SES on the mathematics self-efficacy, anxiety and self-concept factors were estimated simultaneously via standardized partial regression coefficients. MIMIC Model 1 results presented in Table 4.

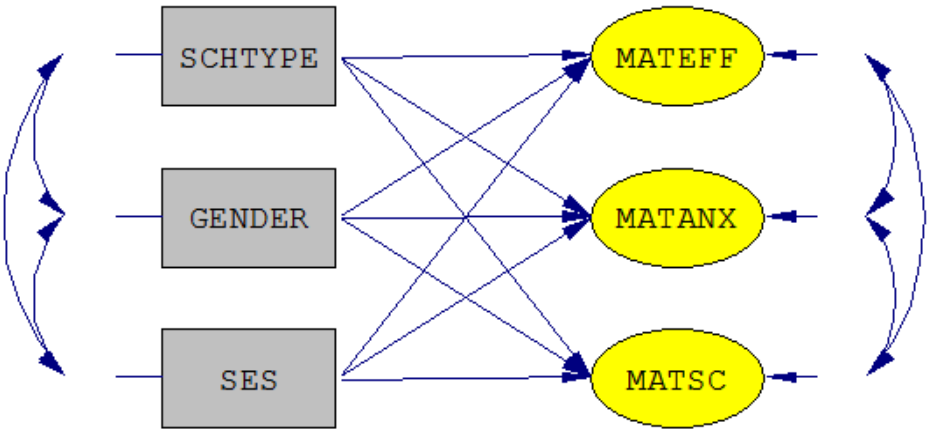


Figure1. MIMIC Model 1

Note. SCHTYPE: school type; GENDER: gender; SES: socioeconomic status; MATEFF: mathematics self-efficacy; MATANX: mathematics anxiety; MATSC: mathematics self-concept.

Table 4. MIMIC Model 1 Results

	Mathematics self-efficacy			Mathematics anxiety			Mathematics self-concept		
	β	*b*	*SE*	β	*b*	*SE*	β	*b*	*SE*
School type	0.23*	0.39	0.07	-0.09*	-0.16	0.07	0.01	0.02	0.07
Gender	0.16*	0.27	0.06	0.01	0.02	0.07	0.09*	0.16	0.07
SES	0.20*	0.15	0.03	-0.10*	-0.07	0.03	0.10*	0.08	0.03

Note. ***p*<0.01. β: standardized path coefficients; *b*: unstandardized estimates; *SE*: standard error of estimates.

The MIMIC Model 1 presented in Figure 1 showed adequate fit ($\chi^2(107)= 395.80$; RMSEA= .059 [.052:.063]; CFI= .97; NFI= .96; GFI=.93). Overall, 16%, 2% and 2% of the variability of mathematics self-efficacy, anxiety and self-concept factors were explained respectively by this MIMIC model.

MIMIC Model 1 confirms that Anatolian high school students were more likely to have higher mathematics self-efficacy and lower anxiety than do general high school students, while other covariates are controlled in the model. However, school type was not associated with students' mathematics self-concept. According to gender, male students were more likely to have higher mathematics self-efficacy and self-concept compared to female students do have. Gender was not associated with mathematics anxiety. Finally, students with higher SES were more likely to have higher mathematics self-efficacy and self-concept and lower mathematics anxiety compared to students do have from lower SES.

In addition to variables in Model 1, mathematics achievement and school mathematics achievement variables were inserted to Model 2. Table 5 shows the path coefficients for the effects of covariates on mathematics self-efficacy, anxiety and self-concept in the MIMIC Model 2.

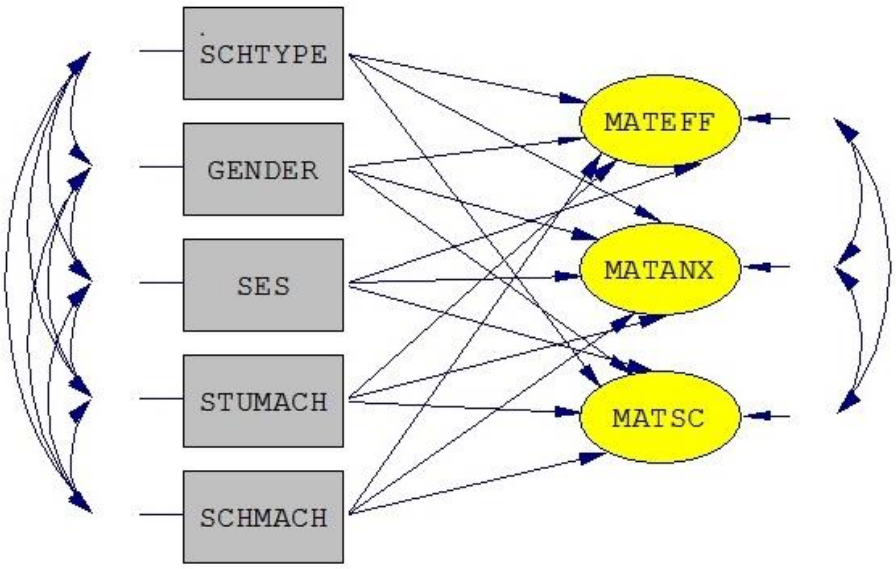


Figure 2. MIMIC Model 2

Note. SCHTYPE: school type; GENDER: gender; SES: socioeconomic status; STUMACH: student mathematics achievement; SCHMACH: school mathematics achievement; MATEFF: mathematics self-efficacy; MATANX: mathematics anxiety; MATSC: mathematics self-concept.

Table 5. MIMIC Model 2 Results

	Mathematics self-efficacy			Mathematics anxiety			Mathematics self-concept		
	β	*b*	*SE*	β	*b*	*SE*	β	*b*	*SE*
School type	-0.04	-0.07	0.10	0.10*	0.19	0.12	-0.09	-0.17	0.11
Gender	0.13*	0.23	0.06	0.05	0.09	0.07	0.06	0.10	0.07
SES	0.11*	0.08	0.03	-0.03	-0.02	0.03	0.06	0.04	0.03
Student mathematics achievement	0.40*	0.35	0.05	-0.47*	-0.44	0.06	0.46*	0.41	0.06
School mathematics achievement	0.07	0.06	0.07	0.12	0.11	0.08	-0.22*	-0.20	0.07

Note. *****p<0.01. β:standardized path coefficients; *b*:unstandardized estimates; *SE*: standard error of estimates.

Model 2 also showed adequate degree of fit to data ($\chi^2(129)= 455.40$*;* RMSEA= .057 [.052:.063]; CFI= .98; NFI=.97; GFI= .93). Overall, 26%, 12%, 10% of the variability of mathematics self-efficacy, anxiety and self-concept factors were explained respectively by MIMIC Model 2.

After controlling the effects of other covariates, the school type variable was examined; Anatolian high school students have higher mathematics anxiety compared to general high school students. Moreover, male students have higher mathematics self-efficacy compared to female students. Similarly, students from higher SES are more likely to have higher mathematics self-efficacy compared to students from lower SES. Student's mathematics achievement was the most significant predictor of the mathematics self-efficacy, anxiety and self-concept factors. Students with high mathematics achievement were more likely to have higher mathematics self-efficacy and self-concept and lower mathematics anxiety compared to students with low mathematics achievement. Finally, school's mathematics achievement was a significant predictor of only mathematics self-concept. Students from schools that have higher mathematics achievement were more likely to have lower mathematics self-concept compared to students from schools that have lower mathematics achievement.

DISCUSSION and CONCLUSION

In this study, the 2012 PISA Turkey data is considered to determine the big fish-little pond effect. The mathematics self-efficacy, self-concept and anxiety factors are examined to explain the relation of each of them with the gender, socioeconomic status, student's mathematics achievement, and school's mathematics achievement. Factor analyses' results support the construct validity of the student questionnaire's mathematics self-efficacy, anxiety and self-concept items. Two MIMIC models were analyzed and the patterns of association with covariates and affective factors were tested simultaneously.

Anatolian high school students have a higher mathematics self-efficacy and lower anxiety than do the general high school students according to the Model 1. However, when the student mathematics achievement and school mathematics achievement variables were inserted to the model, school type was not associated with mathematics self-efficacy. Moreover, Anatolian high school student's mathematics anxiety was higher than that of the general high school students in Model 2, which was opposite to the results of Model 1. Anatolian high school students that are admitted with high mathematics scores compare themselves with the classmates who have similar or higher mathematics scores may induce them to think that they are inadequate. As a result, they might have high mathematics anxiety.

In terms of gender, the male students' mathematics self-efficacy was higher than that of the female students for both models. However, male students' mathematics self-concept was higher than that of the female students at Model 1 whereas at Model 2, gender was not associated with mathematics self-concept.

While investigating the effect of SES, after inserting the variables of student's mathematics achievement and school's mathematics achievement into the Model 2, it was realized that SES was not associated with the mathematics anxiety and self-concept anymore. SES was a significant predictor of mathematics self-efficacy for both models.

The results show that students' mathematics achievement was the most powerful predictor of the mathematics self-efficacy, anxiety and self-concept factors.

Finally, school mathematics achievement was a significant predictor of mathematics self-concept. The identification of increase in school's mathematics achievement yields a decrease in the student's mathematics self-concept may be considered as the most important result of this study. It is straightforward that social comparison at the Anatolian high schools, which have a high school mathematics achievement, has an important role on the students' self-evaluation. Thus, this result provides evidence about not only the Anatolian high school students but also the high achieving school students experience big fish-little pond effect, which is consistent with the international results (Marsh et al., 2014; Marsh & Hau, 2003; Nagengast & Marsh, 2011; Seaton, Marsh, Yeung & Craven, 2011). Besides, with the increase at school's mathematics achievement, there is a tendency for a decrease at student's mathematics self-efficacy and an increase at the anxiety. However, these results were not statistically meaningful as it was in the mathematics self-concept. This can be explained by items that evaluate mathematics self-efficacy having specific cognitive process which causes a higher correlation with the mathematics achievement. Another possible explanation is that students attending to schools that have higher mathematics achievement can evaluate themselves more precisely according to the students attending to the schools that have lower mathematics achievement about which type of mathematics items that they can answer correctly.

By 2014, in Turkey, the general high schools were converted to Anatolian high schools and thus all students were placed according to their high school entrance examination results to decrease the diversity among the quality of high schools and increase overall quality of the general high schools. Consequently, all high school students began to be grouped according to their academic capabilities. When the results of this study is evaluated by considering the new high school entrance system, the student's mathematics achievement is shown to be the most influential variable on the affective

factors; however, especially, the school's mathematics achievement negative impact on the mathematics self-concept is undeniable.

In Turkey, there is a perception about attending a high school that admits students only with high test scores is the most beneficial thing for the students. In fact, rather than being a big fish in a small pond, attending a school with the students of a wide range of ability level and getting a good education might protect particularly the student's mathematics self-efficacy and also their mathematics anxiety and self-concept against the negative impacts of social comparison.

After the PISA study results, there has been a tendency in the world not to group the students according to their academic capabilities. On the contrary, in Turkey, the new high school entrance system makes the diversity of students' academic capabilities decrease among the schools. More studies should be required to discover the possible results of this new high school entrance system.

Author's note: The author gratefully acknowledges support from the TÜBİTAK and London School of Economics and would like to thank Professor Irini Moustaki for her valuable comments.

REFERENCES

Alcı, B., Erden, M., & Baykal, A. (2010). Üniversite öğrencilerinin matematik başarıları ile algıladıkları problem çözme becerileri, özyeterlik algıları, bilişüstü özdüzenleme stratejileri ve ÖSS sayısal puanları arasındaki açıklayıcı ve yordayıcı ilişkiler örüntüsü. *Boğaziçi Üniversitesi Eğitim Dergisi, 25*(2), 53-68.

Altun, F., & Yazıcı, H. (2013). Ergenlerin benlik algılarının yordayıcıları olarak: akademik öz-yeterlik inancı ve akademik başarı. *Kastamonu Eğitim Dergisi, 21*(1), 145-156.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science, 11*(5), 181-185.

Baloglu, M., & Kocak, R. (2006). A multivariate investigation of the differences in mathematics anxiety. *Personality and Individual Differences*, *40*(7), 1325-1335.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117-148.

Bong, M. (1998). Personal factors affecting the generality of academic self-efficacy judgments: Gender, ethnicity, and relative expertise. *Journal of Experimental Education,* 67, 315–331.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270-295. doi:10.1006/ceps.2001.1094

Chiu, M. M., & Xihua, Z. (2008). Family and motivation effects on mathematics achievement: Analyses of students in 41 countries. *Learning and Instruction, 18*, 321-336. doi:110.1016/j.learningstruc.2007.06.003

Cooper, S. E., & Robinson, D. A. (1989). The influence of gender and anxiety on mathematics performance. *Journal of College Student Development, 30*(5), 459-461.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117-140.

Geist, E. (2010). The anti-anxiety curriculum: Combating math anxiety in the classroom. *Journal of Instructional Psychology*, *37*(1), 24-31.

Hammouri, H. (2010). Attitudinal and motivational variables related to mathematics achievement in Jordan: findings from the TIMSS. *Educational Research*, *46*(3), 241-257. doi:10.1080/0013188042000277313

Hattie, J. A. (1992). *Self-concept*. Hillsdale, NJ: Erlbaum.

Hembree, R. (1990). The nature, effects and relief of mathematics anxiety. *Journal for Research in Mathematics Education, 21*(1), 33-46.

Ho, E. S. (2009). Characteristics of East Asian learners: what we learned from PISA. *Educational Research Journal*, *24*(2), 327-348.

Ho, H., Senturk, D., Lam, A. G., Zimmer, J. M., Hong, S., Okamoto, Y., Chiu, S., Nakazawa, Y., & Wang, C. (2000). The affective and cognitive dimensions of math anxiety: A crossnational study. *Journal for Research in Mathematics Education, 31*(3), 362-379.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

İş Güzel, Ç., & Berberoğlu, G. (2010). Students' affective characteristics and their relation to mathematical literacy measures in the Programme for International Student Assessment (PISA) 2003. *Eurasian Journal of Educational Research*, *4*, 93-113.

Jain, S., & Dowson, M. (2009). Mathematics anxiety as a function of multidimensional self-regulation and self-efficacy. *Contemporary Educational Psychology*, *34*(3), 240-249.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software International. Inc.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80 for Windows*. Lincolnwood, IL: Scientific Software International, Inc.

Kurbanoğlu, N. I., & Takunyacı, M. (2012). An investigation of the attitudes, anxieties and self-efficacy beliefs towards mathematics lessons high school students' in terms of gender, types of school, and students' grades. *International Journal of Human Sciences*, *9*(1), 110-130.

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30*(5), 520-540.

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J., Abdelfattah, F., & Nagengast, B. (2014). The big-fish-little-pond effect in mathematics a cross-cultural comparison of US and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology, 45*(5), 777-804.

Marsh, H. W., Chessor, D., Craven, R., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal*, *32*(2), 285-319.

Marsh, H. W., & Hau, K. T. (2003). Big-Fish--Little-Pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, *58*(5), 364-376.

Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, *6*(4), 311-360.

Marsh, H. W., Kong, C. K., & Hau, K. T. (2000). Longitudinal multilevel models of the big-fish-little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology*, *78*(2), 337-349.

Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, *38*(2), 321-350.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397-416.

Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish-little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, *44*(3), 631-669.

McCarthy, J. M., & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of Selection and Assessment, 13*(4), 282-295.

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the UK, UK countries and the world: The big-fish-little-pond-effect for PISA 2006. *Educational Psychology, 31*(5), 629-656. doi: 10.1080/01443420.2011.586416

OECD (2009). *PISA data analysis manual SPSS (2nd edition)*. Paris: Publications.

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en

Özel, M., Çağlak, S., & Erdoğan, M. (2012). Are affective factors good predictor of science achievement? Examining the role of affective factors based on PISA 2006. *Learning and Individual Differences, 24,* 73-82. doi:10.1016/j.lindif.2012.09.006

Özgen, K., & Bindak, R. (2011). Lise öğrencilerinin matematik okuryazarlığına yönelik öz-yeterlik inançlarının belirlenmesi. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*(2), 1073-1089.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543-578.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, *24*(2), 124-139.

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*(2), 193-203.

Reyes, L. H. (1984). Affective variables and mathematics education. *The Elementary School Journal, 84*(5), 558-581.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling.* Routledge NY.

Seaton, M., Marsh, H. W., Yeung, A. S., & Craven, R. (2011). The big fish down under: Examining moderators of the 'big-fish-little-pond'effect for Australia's high achievers. *Australian Journal of Education, 55*(2), 93-114.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*, 27-41.

Taşdemir, C. (2012). Lise son sınıf öğrencilerinin matematik öz-yeterlik düzeylerinin bazı değişkenler açısından incelenmesi (Bitlis ili örneği). *Karadeniz Fen Bilimleri Dergisi, 2*(6), 39-50.

Wigfield, A., & Meece, J. L. (1988). Math anxiety in elementary and secondary school students. *Journal of Educational Psychology*, *80*(2), 210-216.

Wilkins, J. L. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education, 72*(4), 331-346.

Wilkins, J. L. M., Zembylas, M., & Travers, K. J. (2002). Investigating correlates of mathematics and science literacy in the final year of secondary school. In D. F Robataille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data*. Boston, MA: Kluwer Academic.

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin*, *106*(2), 231-248.

Wu, M. L., & Adams, R. J. (2002). *Plausible values–why are they important.* Paper presented at the International Objective Measurement Workshop, New Orleans, LA.

Zeidner, M., & Schleyer, E. J. (1999). The big-fish–little-pond effect for academic self-concept, test anxiety, and school grades in gifted children. *Contemporary Educational Psychology*, *24*(4), 305-329.

UZUN ÖZET

Giriş

Sosyal karşılaştırma teorisine göre, objektif standartların bulunmadığı durumlarda, insanlar çevrelerinde bulunan diğer insanları temel alarak öz-değerlendirme yapmaktadır (Festinger, 1954). Okullar bu anlamda sosyal karşılaştırmanın en fazla yapıldığı ortamlar olarak karşımıza çıkmaktadır. Akademik yeterliklerine göre gruplandırılarak eğitim gören öğrencilerler, farklı yetenek seviyesindeki öğrencilerle beraber eğitim görenlerden daha düşük benlik algısına sahip olmaktadır (Nagengast ve Marsh, 2011). Büyük balık-küçük göl etkisi olarak adlandırılan bu durum özellikle öğrencilerin başarılarına göre sınıflandırıldığı eğitim sistemlerinde görülmektedir (Ho, 2009).

Türkiye'de öğrenciler akademik başarılarına göre gruplandığından, büyük balık-küçük göl etkisinin okul seviyesinde yaşandığı ve öğrencilerde oluşan farkındalığın, öğrencilerin matematik öz-yeterlik, kaygı ve benlik algılarını etkileyebileceği düşünülmektedir.

Bu çalışmanın amacı, Uluslararası Öğrenci Değerlendirme Programı (Programme for International Student Assessment, [PISA]) 2012 Türkiye verisi üzerinde Anadolu lisesi öğrencileriyle genel lise öğrencilerinin; cinsiyet, sosyoekonomik statü, öğrenci matematik başarısı, okul matematik başarısı değişkenlerinin matematik dersine yönelik öz-yeterlik, benlik ve kaygı değişkenleri arasındaki ilişkiyi açıklayarak büyük balık-küçük göl etkisini belirlemeye çalışmaktır.

Yöntem

Çalışma, PISA 2012 Türkiye verisinin bir yapısal eşitlik modellemesi olan MIMIC (Multiple Indicator Multiple Cause-Çoklu Gösterge ve Çoklu Neden) model ile analiz edilmesiyle gerçekleştirilmiştir. MIMIC modelde, her bir neden değişkeninin özgün etkisi diğer neden değişkenlerinin etkisi arındırılarak hesaplanabilmektedir.

Bu çalışmadaki veriler PISA 2012 uygulamasından elde edilmiştir. Çalışmada, öğrenci anketindeki okul türü, cinsiyet, sosyoekonomik statü, matematik öz-yeterlik, benlik ve kaygı ile ilgili maddeler kullanılmıştır. Türkiye verisinde matematik öz-yeterlik, kaygı ve benlik maddelerini cevaplayan, 36'sı Anadolu ve 52'si genel lise olmak üzere toplam 88 farklı liseden 771 öğrenci örnekleme alınmıştır.

Öncelikle açımlayıcı faktör analiziyle matematik dersine yönelik öz-yeterlik, kaygı ve benlik maddelerinin faktör yükleri belirlenmiş ve bu değişkenlere ait maddelerle doğrulayıcı faktör analizi yapılmıştır. Ardından MIMIC yapısal eşitlik modeli test edilerek Anadolu ve genel liseler için yol

katsayıları karşılaştırılmıştır. İki MIMIC modeli analiz edilmiş ve bu analizlerle neden değişkenleri ve duyuşsal faktörler arasındaki örüntü eş zamanlı olarak test edilmiştir.

Model uyumunun değerlendirilmesinde RMSEA (Root Mean Squared Error of Approximation-Ortalama Karekök Hata Tahmini), CFI (Comparative Fit Index-Karşılaştırmalı Uyum İndeksi), NFI (Normed Fit Index-Normlaştırılmış Uyum İndeksi) ve GFI (Goodness of Fit Index-Uyum İyiliği İndeksi) uyum indeksleri kullanılmıştır. Analizler LISREL 8.8 (Jöreskog ve Sörbom, 2006) programı SIMPLIS komut dili kullanılarak gerçekleştirilmiştir.

Anketin ilgili maddelerinin öz değeri 1'den büyük üç faktörlü bir yapıya sahip olduğu belirlenmiş ve maddeler faktörler altında kümelenmelerine göre matematik benlik, öz-yeterlik ve kaygı olarak adlandırılmıştır. Bu üç faktör toplamda varyansın %58,62'sini açıklamaktadır. Matematik benlik faktörüne dört madde (α=0,86), matematik öz-yeterlik faktörüne altı madde (α=0,74) ve matematik kaygı faktörüne dört madde yüklenmiştir (α=0,82). İkinci adımda, açımlayıcı faktör analizinden elde edilen sonuçlar, doğrulayıcı faktör analiziyle sınanmıştır. Doğrulayıcı faktör analiziyle ortaya çıkan üç boyutlu yapı kabul edilebilir bir model olup açımlayıcı faktör analiziyle ortaya çıkan yapının doğrulandığı söylenebilir. Üçüncü adımda, ölçme modeli doğrulayıcı faktör analizi ile sınandıktan sonra modele yapısal kısım ilave edilerek MIMIC model elde edilmiş ve parametreler kestirilmiştir.

MIMIC Model 1 ve 2 için hesaplanan veri uyumu kabul edilebilir seviyededir. Toplamda MIMIC Model 1 tarafından matematik öz-yeterlik, kaygı ve benlik faktörlerindeki değişkenliğin sırasıyla %16'sı, %2'si ve %2'si açıklanmıştır. MIMIC Model 2'de, Model 1'deki değişkenlere ilaveten öğrenci matematik başarısı ve okul matematik başarı ortalaması değişkenleri de modele eklenmiştir. MIMIC Model 2 tarafından matematik öz-yeterlik, kaygı ve benlik faktörlerindeki değişkenliğin sırasıyla %26'sı, %12'si ve %10'u açıklanmıştır.

Sonuç ve Tartışma

Bu çalışmada, 2012 PISA Türkiye verisi kullanılarak büyük balık-küçük göl etkisini belirlemek amacıyla, matematik öz-yeterlik, benlik ve kaygı faktörlerinin; cinsiyet, sosyoekonomik statü, öğrenci matematik başarısı ve okul matematik başarısı değişkenleriyle olan ilişkileri incelenmiştir.

Model 1'e göre Anadolu lisesi öğrencilerinin matematik dersine yönelik öz-yeterlikleri genel lise öğrencilerininkinden daha yüksek ve matematik kaygıları daha düşüktür. Bununla birlikte, modele öğrenci matematik başarısı ve okul matematik başarısı değişkenleri ilave edildiğinde okul türü matematik öz-yeterliği ile ilişkili değildir. Model 2'de Anadolu lisesi öğrencilerinin matematik kaygıları genel lise öğrencilerininkinden daha yüksektir. Bu sonuçlar Model 1'den elde edilen sonuçların tersidir. Yüksek matematik puanları ile Anadolu liselerine kabul edilen öğrencilerin, kendilerini benzer veya daha yüksek matematik başarısına sahip sınıf arkadaşlarıyla karşılaştırıyor olmaları bu öğrencilerin kendilerini yetersiz görmelerine sebep olmuş olabilir. Anadolu lisesi öğrencilerinin yüksek matematik kaygısı bu durumun bir sonucu olabilir.

Cinsiyet değişkeni incelendiğinde, her iki modelde de erkek öğrencilerin öz-yeterlik algılarının kız öğrencilerinkinden daha yüksek olduğu belirlenmiştir. Sadece, Model 1 için erkek öğrencilerin matematik benliği kız öğrencilerinkinden daha yüksek iken, Model 2'de cinsiyet ile matematik benlik algısı ilişkili değildir. Her iki model için de sosyoekonomik statü, matematik öz-yeterlik değişkeninin istatistiksel olarak anlamlı bir tahmin edicisidir. Sonuçlar göstermektedir ki öğrenci matematik başarısı, matematik öz-yeterlik, benlik ve kaygı değişkenlerinin en güçlü tahmin edicisidir. Son olarak, okul matematik başarısı matematik benlik algısının istatistiksel olarak anlamlı bir tahmin edicisidir. Bu çalışmanın en önemli bulgusu, okul matematik başarı ortalaması arttıkça öğrencilerin matematik benlik algılarının düşme eğiliminde olduğunun belirlenmesidir. Okul matematik başarı ortalamasının yüksek olduğu Anadolu liselerinde sosyal karşılaştırmanın öğrencilerin öz-değerlendirmesinde önemli bir rol oynadığı aşikardır. Bu durum, büyük balık-küçük göl etkisinin sadece Anadolu lisesi öğrencileri tarafından değil, başarı ortalamasının yüksek olduğu liselere devam eden öğrenciler tarafından da yaşandığına dair kanıt sunmaktadır. Okulun matematik başarısı arttıkça öğrencilerin matematik öz-yeterlikleri düşme ve kaygıları artma eğiliminde olmakla

beraber sonuçlar matematik benlik algısında olduğu gibi istatistiksel olarak anlamlı bir değişime işaret etmemektedir.

Türkiye'de 2014 yılı itibarıyla, liseler arasındaki kalite farklılıkların azaltılması ve genel liselerde verilen eğitimin seviyesinin artırılması amacıyla genel liseler, Anadolu liselerine dönüştürülmüş ve böylece tüm öğrenciler liselere giriş sınav sonuçlarına göre okullara yerleştirilmiştir. Sonuç olarak, öğrenciler yetenek seviyelerine göre bir arada olacak şekilde ayrıştırılmaya başlanmıştır. Türkiye'de öğrencilerin yüksek puanla öğrenci kabul eden bir okula devam etmesinin onlar için en iyisi olduğuna dair genel bir algı oluşmuştur. Aslında sınavla öğrenci kabul eden bir lisede küçük gölde-büyük balık olmak yerine farklı yetenek seviyesindeki öğrencilerin birarada eğitim görmeleri, sosyal karşılaştırmanın öğrencilerin matematik öz-yeterlik algısı başta olmak üzere matematik kaygısı ve benlik algısı üzerindeki olumsuz etkisinden de koruyabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1),15-33

Exploring Variability Sources in Student Evaluation of Teaching via Many-Facet Rasch Model*

Ders Değerlendirme Anketinde Varyans Kaynaklarının Çok Yüzeyli Rasch Modeliyle Değerlendirilmesi

Bengü BÖRKAN **

Abstract
Evaluating quality of teaching is important in nearly every higher education institute. The most common way of assessing teaching effectiveness takes place through students. Student Evaluation of Teaching (SET) is used to gather information about students' experiences with a course and instructor's performance at some point of semester. SET can be considered as a type of rater mediated performance assessment where students are the raters and instructors are the examinees. When performance assessment becomes a rater mediated assessment process, extra measures need to be taken into consideration in order to create more reliable and fair assessment practices. The study has two main purposes; (a) to examine the extent to which the facets (instructor, student, and rating items) contribute to instructors' score variance and (b) to examine the students' judging behavior in order to detect any potential source of bias in student evaluation of teaching by using the Many-Facet Rasch Model. The data set includes one thousand 235 students' responses from 254 courses. The results show that a) students greatly differ in the severity while rating instructors, b) students were fairly consistent in their ratings, c) students as a group and individual level are tend to display halo effect in their ratings, d) students are clustered at the highest two categories of the scale and e) the variation in item measures is fairly low. The findings have practical implications for the SET practices by improving the psychometric quality of measurement.

Keywords: Student evaluation of teaching, Many Facet Rasch Model, psychometric analysis

Öz
Öğretim niteliğini değerlendirmek neredeyse her yükseköğretim kurumlarında önemlidir. Öğretimin etkinliğini değerlendirmenin en yaygın yöntemi öğrenciler üzerinden gerçekleştirilmektedir. Ders değerlendirme anketi (DDA) yoluyla dönemin herhangi bir zamanında öğrencilerin ders ve öğretim elemanı hakkındaki görüş ve tecrübelerine ilişkin bilgi toplanır. DDA, puanlayıcı aracılı bir tür performans değerlendirmesi olarak kabul edilebilir. Bu kez öğrenciler puanlayıcı, öğretim elemanları ise değerlendirilendir. Performans değerlendirme, bir puanlayıcı aracılı değerlendirme süreci olduğunda, ekstra önlemler, daha güvenilir ve adil bir değerlendirme uygulamaları oluşturmak için dikkate alınması gerekir. Bu çalışmanın amacı, a) öğretim elemanlarının puanlarındaki farklılığa/varyansa, puanlama sürecindeki yüzeylerin (öğretim elemanı, öğrenci ve değerlendirme maddeleri) ne derece katkı sağladığını ve b) öğrencilerin yaptiklari puanlamalarda yanlılığa yol açacak potansiyel kaynakları çok yüzeyli Rasch modeli yardımıyla incelemektir. Çalışmada kullanılan veri seti 254 dersten 1.235 öğrencinin değerlendirmelerini kapsamaktadır. Sonuçlara göre a) öğrenciler, öğretim elemanlarını değerlendirirken farklı katılık/cömertlik derecesi göstermektedirler, b) çoğu öğrenci kendi içinde oldukça tutarlı, c) grup olarak, değerlendirmelerde halo etkisi oldugu görülmektedir, d) öğrenciler beşli ölçeğin üst puanlarında kümelenmişlerdir, e) madde zorluk değerlenlerindeki varyasyon çok düşüktür. Bu bulguların, DDA'nin psikometrik özelliklerinin daha nitelikli hala getirilmesi yönünde sonuçları vardır ve bunlar makalede tartışılmıştır.

Anahtar Kelime: Ders değerlendirme anketi, Çok Yüzeyli Rasch Modeli, psikometrik analiz

* Preliminary results of this study was presented at the European Conference on Educational Research 2015
** Yard. Doç. Dr., Boğaziçi University, Istanbul-Turkey, e-posta: bengu.borkan@boun.edu.tr

DOI: 10.21031/epod.298462
Geliş Tarihi: 29.09.2016
Kabul Tarihi: 28.02.2017

INTRODUCTION

Student Evaluation of Teaching (SET) which is used to gather information regarding students' experiences with a course and instructor's performance at some point of semester seems to be to the most common ways of gathering data for supposedly both formative and summative evaluation purposes (Gravestock, & Gregor-Greenleaf, 2008; Penny, 2003; Seldin, 1993; Zabaleta, 2007). By means of SET results, administrators in higher education institutes aim to (a) improve teaching quality, (b) provide input for appraisal exercises (e.g., tenure/promotion decisions), and (c) provide evidence for institutional accountability (Seldin, 1993; Spooren, Brock & Mortelmans, 2013).

SET can be considered a type of performance assessment. In performance assessment, a person (examinee) displays performance and/or construct a product and, quality of this performance or product is evaluated by at least one evaluator/rater. When performance assessment becomes a rater mediated assessment process, extra measures need to be taken into consideration in order to create more reliable and fair assessment practices. One of the most common threat in rater mediated assessment is called 'rater variability'. This term generally describes the variability that is linked to rater characteristics (i.e. lenient, severe, gender), not to the performance of person being evaluated (Eckes, 2009). In other words, rater variability threatens the validity and fairness of performance assessment when measurement is involved construct irrelevant variance in examinee scores (Lane & Stone, 2006; Messick, 1998). Eckes states that "This long, and possibly fragile, interpretation–evaluation–scoring chain highlights the need for carefully investigation of the psychometric quality of rater-mediated assessments. One of the major difficulties facing the researcher, and the practitioner alike, is the occurrence of rater variability." (p.4).

As literature point out, both theoretical and psychometric issues remain unresolved for SET questionnaires (Gravestock & Gregor-Greenleaf, 2008). Studies have been accumulated around two main concerns which are in fact relevant to each other. The first concerns focus on the question whether the score obtained from students' evaluations are valid and actually measure what we intent to measure so called teaching effectiveness. The second concern focus on potential bias sources which treats the reliability and validity of our measures (Gursoy & Umbreit 2005) such as gender of the instructor, expected grade. The purpose of this study is to (a) examine the extent to which the facets (instructor, student, and rating items) contribute to instructors' score variance and (b) examine the students' judging behavior using the Many-Facet Rasch Model in order to detect any potential source of bias in student evaluation of teaching.

Evaluating Quality of Teaching in Higher Education

Because of the great extent use of SET, an enormous literature has been collected since early 1920 in which the first SET was administered at the University of Washington (Seldin, 1993; Zabaleta, 2007). Since then, some issues such as validity of SET has remained, and other issues like the use of SET results to improve teaching, has recently come to researchers' attention. Majority of the research has been conducted in North American, Australian and UK teaching context (Gravestock, & Gregor-Greenleaf, 2008; Zabaleta, 2007). Majority of those studies have generally positive position for the use of SET (such as Abrami, 2001; Beran, Violato, & Kline, 2007; Gravestock, & Gregor-Greenleaf, 2008; Marsh, 1987); on the other hand, some have displayed a skeptical attitude toward the use of SET since SET could produce bias results due to teacher and course characteristics which are believed to be irrelevant with the quality of teaching (Dodeen, 2013; Koh & Tan, 1997; Williams & Ceci, 1997). Literature displays ambivalent research findings regarding the validity of this method and the use of its results. While some studies claim that SET provides valid data in general (e.g. Marsh, 1984, Nelson & Lynch, 1984, Zangenehzadeh, 1988) and no bias in particular, other studies reported bias in the data and concluded deficiency in validity of SET scores (e.g. Centra, 1993, Haladayna & Hess, 1994, Marsh, 1987, Marsh & Roche, 2000). Therefore, it is suggested that the results of SET should not be used alone for high stake decision such as retention, promotion or tenure (Penny, 2003).

Economic and political changes in the World have been pushing higher education institutes to exhibit their performances equally well in both research and teaching arena. Moore and Kuol (2005) argued that SET provides us quantitative data that we can use for comparison is imprecise ways of evaluating and comparing teaching effectiveness. Therefore, educational institutions should be well informed about how to present, interpret and use these sorts of data (Gravestock, & Gregor-Greenleaf, 2008). Although it is widely accepted that SET should not be the only toll to evaluate ones teaching quality, SET result will continue to be used for longer time as performance indicator of teaching effectiveness (Penny, 2003).

We have big assumption based on the idea that data obtained SET questionnaire is a good measure of teacher effectiveness which leads to students' learning and better education. Well informed decisions should be based on a vigorous SET system including valid and reliable data collection instrument and dependable data collection procedure. Many critical studies highlighted the weakness of student evaluations of teaching and questioned its validity. Instructor's sex or personality could be a determining factor in students' rating. Being a female instructor can be disadvantageous (Basow & Martin, 2012) or perceived attractiveness/expressiveness of the instructor is a shaping factor in students' judgment on the quality of teaching; (Cashin, 1995). Marsh (1982) reported that SET appears to be subject to substantial halo effects, which means students answers the different dimension of the instrument in a similar way. Barnes and colleagues (2008) and Wachtel (1998) reported students' ratings on different dimension were significantly correlated with students' expected course grade. Students tend to rate the instructor or instruction more highly in smaller courses (Hoyt, Che, Pallett, Gross, 1999) and age is negatively correlated with evaluation scores (Zabaleta, 2007). Grading leniency/severity can bias student ratings (Basow & Martin, 2002).

Student Evaluations of Teaching Questionnaire

Effective teaching is considered as a construct that we use to explain desired instructor/teachers behavior in educational process. None of the construct can be directly observed and measured and, therefore we need an operational definition for them. In other words, we need a list of related behaviors that are associated with our construct. Unfortunately, no consensus on the definition of what effective teachings is in this sense. Ory and Ryan (2001) argue that there is no "universal set of characteristics of effective teachers and courses that should be used as a target…"(p.32). Therefore, various measurement tools are available; almost every institution developed their own questionnaire by considering institutional needs, climates and priorities. Keeley (2012) called this questionnaires "home grown", I called "tailor made".

Existing questionnaires have different content/items in various lengths. SET consists of a questionnaire which usually includes mixture of open-ended (qualitative data) and closed-ended items (quantitative data) with a rating scale. Items are usually related with different dimensions of teacher effectiveness such as planning, organization, grading, interaction, instruction, learning, fairness of grading. By means of including all different dimension of teaching better content and construct validity could be achieved. However, this makes the questionnaire longer.

A number of researchers conducted a study for the purpose of identifying the dimensions, sub-construct or factors of the construct that is usually named as students' perceptions of teaching effectiveness. Here some examples for a well-developed, psychometrically evaluated instrument. Barnes and colleagues (2008) developed a questionnaire with 14 items. They identified two distinct factor; teaching excellence and teaching readiness. In another study, Mortelmans and Spooren (2009) developed a questionnaire including 37 items and 12 dimensions of effective teaching; build-up of subject matter, Linking-up with foreknowledge\ content validity of examination, presentation skills, value of subject matter, course difficulty, harmony organization course learning\ course materials, clarity of course objectives, help of teacher during learning process, formative evaluation and authenticity of the examination. Marks (2000) explore five underlying constructs for his student evaluation questionnaire: organization, expected/fairness of grading, workload/difficulty, perceived and instructor liking/concern, learning. Marsh (1982) reported nine dimensions of teaching

effectiveness with 35 items; learning/value, organization, enthusiasm, breadth of coverage, group interaction, individual rapport, examinations and grading, assignments, and workload. The questionnaire has very high internal consistency coefficient and it produced stable results over time. Ginns, Prosser and Barrie (2007) developed their own 23 item five factor SET questionnaire. They named factors as generic skills scale, appropriate assessment, good teaching, appropriate workload and clear goals and standards. They report high internal consistency and inter-rater agreement.

In some higher education institutions SET may play an important role and it effects teaching climate of institution in particular. Negative attitude of instructor towards SET were mention in the literature (e.g. Spooren et al, 2013). Given that instructors are the primary users of this system, their trust is very curial to fully utilize the potential of SET. Primary reason of instructor not to trust is the belief of potential bias in student rating. Yet, negative findings regarding the validity mentioned above elevate their concern. Consequently, SET needs to come under the spotlight in order to develop instructor trust and to increase the practical usefulness of SET.

Rater-Mediated Performance Assessment

The score of examinee on a performance task depends on not only examinee ability but also other various facets related to the nature of assessment. Three most commonly seen aspects (facets) are the ability of the examinee and the difficulty of the performance task (Mulqueen , Baker & Dismukes, 2002) and the rater effect. Rater effects in an evaluation process appears in different forms such as halo effect, rater severity/leniency or central tendency (Hoyt, 2000; Myford & Wolfe, 2003). Such rater effects introduce a method variance for scores. Previous research in different settings show that significant rater effects exist in rater mediated performance assessment (Eckes, 2005). For example, The meta-analysis study shows that 37% of examinee performance can be explained with rater effect and rater-ratee interaction (Hoyt & Kerns, 1999). Different procedures can be used to control reliability of scoring in evaluation processed where there are multiple evaluators.

A Many-Facet Rasch Model approach

Like G Theory, a Many-Facet Rasch Model (MFRM) approach allows researchers or practitioners to analyses potential sources of errors in rating processes. A MFRM developed by Linacre (1989) is based on the basic Rasch model (Rasch, 1960/1980). The Rasch model, a one-parameter latent trait model, provides item free estimates of each person ability and person free estimates of item difficulty and places both estimates on an equal-interval log-linear scale (Wright & Stone, 1979). In other words, estimates of person measures are independent of the difficulty of item or task in measurement processes, and estimates of item or tasks measures are independent of the specific group of people ability (Sudweeks, et al, 2005).

Since a MFRM is a member of the Rasch family, it possesses all of the characteristics of a basic Rasch model and more. MFRM allows assessment of various variability sources in the rating score, for instance examinee ability, task difficulty, rater severity and interaction of these facets. In short, a MFRM has the following benefits;

a) If the data fits the model, each facet are estimated independently (Linacre & Wright, 2002); in other words, the measures obtained by the model are sample, item, and condition-free. Therefore, function of facet can be evaluated separately (Myford & Wolfe, 2003).
b) Since person estimates, item estimates and all other facets' estimates are located on the same logit scale, comparisons between facets are possible.
c) Individual level effect (besides group level effect given in previous Bullet b) within in each facet are examined more closely; for instance, which raters rate more severely or which raters disagree other raters.
d) The MFRM provides us goodness-of-fit statistics showing degree of data fit to the model, and they help us to interpret the fit of each single element in each facet (Sudweeks et al, 2005).

e) The MFRM provides bias analysis, that is, the analysis of the interactions between elements of different facets (see Linacre, 2009a, for details). For instance, researchers can examine raters' severity depends upon the characteristics of ratee or the condition of the ratings (Myford & Wolfe, 2003).

Many Facet Rasch Model

Many facet Rasch model extends the Rasch model into more complex situation including more than two facets (i.e. examinee and item) of interest of interest (Linacre, 1989). This model is particularly useful for analysis of subjectively rated performance assessment and/or various tasks in different difficulty level:

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \tau_k \qquad \text{(MFRM, 6)}$$

where

P_{njimk}, = probability of person *n* receiving a rating of *k* on criterion *i* from rater *j*,

$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,

β_n = ability of person *n*,

δ_i = difficulty of criterion *i*,

γ_j = severity of rater *j*,

τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

MFRM is an additive model and can be expanded to as many facets as we like. Besides persons and item facets, other facets that are susceptible to contributing construct irrelevant variances in measurement, such as raters, occasion, and task facets, can be added to the model. As in Rasch model, this model calibrates each facet on a common logit linear scale after raw scores are corrected for inconsistencies among raters' severity, differences in the relative task difficulty (Lunz, Wright & Linacre, 1990). Along the logic scale, the higher the number is, the more lenient the rater is; the more negative the number is, the more severe the rater is. Moreover, MFRM allows us to detect unusual interactions called as bias between raters and tasks/items, or raters and particular examinee (Linacre, 1994).

METHODOLOGY

Data Source

This study will utilize student evaluation of teaching data collected in the undergraduate courses of a mid-size university in a big city. This public university is located on the north western part of Turkey and serving around 11 thousand students in 32 undergraduate programs and four thousand graduate students in 56 master and 32 doctoral programs.

The university has a 150 year-long historical period. From the beginning, significance of teaching excellence has traditionally been emphasized. The university first started to administer paper based student evaluation of teaching questionnaire at the end of every semester. In 2008, instructors of some courses became a volunteer for web based version of SET questionnaire. For those courses, student filled out online questionnaires. Until 2010 Fall semester, the mixed method administration for student evaluation of teaching questionnaire had been continued; online and paper-based. Since 2010 Fall, instructors of all graduate and undergraduate courses in the university have been evaluated by students online.

SET questionnaire has three parts (see the appendix for the content of the item). The first part includes five items about a course and 10 items about instructor effectiveness. Each item has a five point rating scale (5: Excellent, 1: Poor). The second part includes several items about courses

related information such as students' attendance to the course, expected grade from the course, whether the course is required or elective in students' program. The last part includes a textbox where students may write any comments about the course and the instructor.

Sample

In the current study the analysis requires connectedness in data; when every judge rates every person in a study, data is complete. However, if people were rated by different judges and judges cannot be linked through people, we would have subsets in the data and then, connectedness becomes a problem. Therefore, only one faculty out of five was purposefully chosen to guarantee the connectedness; Since Faculty of Science and Art offers courses to all students at the university, selected students in the data set have high chance to provide representative student sample for the population.

In 2015 Fall Academic semester, response rates varied in undergraduate courses. I only included courses if more than six students' evaluated the instructor of the course in order to secure the connectedness among courses. After data cleaning, 254 courses and 1235 students were left in the data.

MFRM Analysis

Before Rasch analysis was conducted, the assumption of unidimensionality was checked. First, factorial structure of the scale was examined by using both exploratory with IBM Statistical Package for Social Sciences 21 and confirmatory factor analysis with MPlus.

Rasch analysis was completed using Facets v. 3.71.4 (Linacre, 1987-2014). I adopted Rasch Rating Scale Models (Andrich, 1978); a three facet and four facet Rasch models. Students, instructor and item are the common facets in both models. A course type and expected grade is an additional facet in the four facet model. Three mathematical models are given in the Appendix. Facets reported that subset connectedness was obtained in the data.

Listed below are some important indexes and evaluation criteria for the analyses of this research.

a) The Infit and Outfit mean square (MnSq) statistics: The Infit and Outfit MnSq statistics reflect the discrepancy between observed and model-driven expected responses and flag unexpectedness in the data (Linacre and Wright, 2002). The value of these statistics range from zero to infinity. In case of perfect correspondence these values become one. A value greater than one indicates that variance is higher than expected. Regarding rater fit statistics, high variance means that a rater rate inconsistently and unpredictably. A value below one signals the existence of lower variance in the data than that predicted by the model. In the case of rater facet, these statistics can be interpreted as too predictable rater behavior. The rater either rates too consistent or do not distinguish between different performances. Linacre and Wright suggest that the Infit and Outfit MnSq statistics values between 0.5 and 1.5.
b) The Separation Ratio (G): G represents a measure of the spread of the estimates relative to their measurement error. It ranges from one to infinity. G = 2, for instance, means that the dispersion in the measures of the elements in the facet is two times greater than the imprecision in their estimates (Wright, 1996). While high G value is desired for item and person facet, low G value is desired for rater facet.
c) The reliability of Separation Index (R): R shows how reproducibly different the measures are. It ranges between zero and one. If R is close to one, there is a high probability that the elements of the facet with high measure estimates actually have higher measures than those with low measure estimates (Linacre, 2009). Similar to G value, while high G value is desired for item and person facet, low G value is desired for rater facet.

f) The Fixed (all-same) chi-square statistics: Hypothesis test is conducted to determine whether or not the estimations of each elements of a facet have the same estimates after accounting for measurement error.
g) Bias analysis (interaction): The interaction between facets will be evaluated by using z-score. An absolute value of z-score greater than 2.0 is considered as an indicator of statistically significant interaction between facets.

RESULTS

Assumption of Unidimensionality

Internal structure of the scale was investigated with exploratory factor analysis (EFA) first and then, confirmatory factor analysis (CFA). While EFA conducted on SPSS yielded one factor structure, CFA conducted on Mplus confirmed one factor model. Moreover, item fit values (examine in details later) show that the data fit to the Rasch model is acceptable and therefore assumption of unidimensionality was secured. Out of 63,810 data points, 614 (0.94%) have a standardized residuals bigger or smaller than three, 2.870 (4.49%) have standardized residuals bigger or smaller than two. These numbers shows that data model fit is acceptable.

The Rating Scale

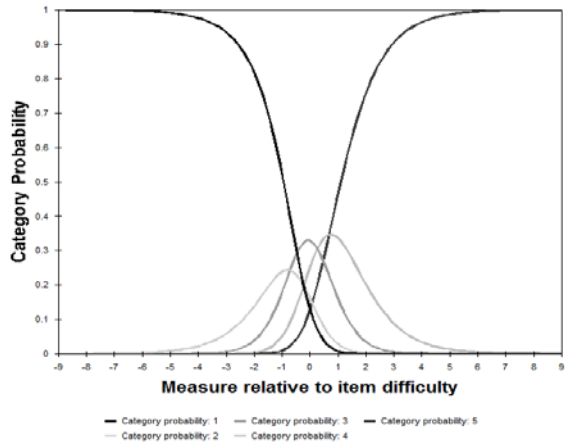


Figure 1. Probability Curves of Five Categories in the Scale

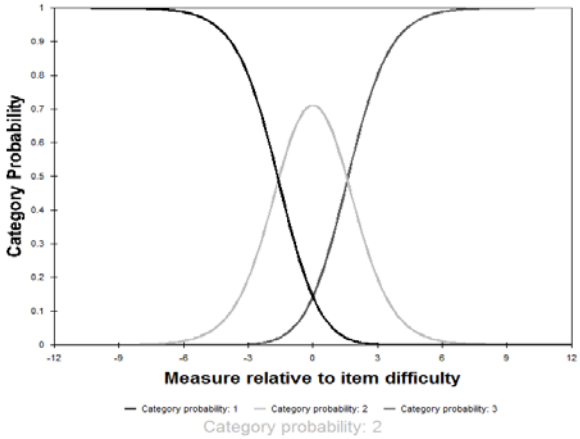


Figure 2. Probability Curves of Three Categories in the Scale

The response scale for items has five categories. This scale is evaluated by how well every point category in the scale conforms to expectations; Figure 1 shows probability curves of categories in the scale. Overlapping categories indicates that the distinction between rating categories students is not clear to students. The measures for thresholds for five categories scale given in Figure 3 show that disordered thresholds exist. This shows that the five category scale does not function as we wish. It is seen that students tend to choice either the first or the last category in the scale.

A five category scale

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	Obsd-Expd	Response
Category	Counts		Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	Diagnostic	Category
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	Residual	Name
1	5496	5091	8%	8%	-.33	-.36	1.3			(-1.74)		low	low	100%	-1.1	lowest
2	4410	4410	7%	16%	.05	.01	1.5	-.03	.02	-.69	-1.22		-.80	23%	-.6	
3	9332	9332	15%	31%	.35	.37	1.0	-.57	.01	-.03	-.34	-.30	-.33	29%	-.5	middle
4	12592	12592	21%	52%	.74	.81	.9	.28	.01	.65	.28	.28	.22	29%		
5	31980	28905	48%	100%	1.59	1.57	1.0	.32	.01	(1.85)	1.26	.32	.86	100%	2.0	highest
										(Mean)		(Modal)	(Median)			

A three category scale

DATA				QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	Obsd-Expd	Response
Category	Counts		Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	Diagnostic	Category
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	Residual	Name
1	5496	5091	8%	8%	-.61	-.69	1.2			(-2.69)		low	low	100%	-.6	lowest
2	26334	26334	44%	52%	.77	.80	1.2	-1.60	.02	.00	-1.68	-1.60	-1.63	71%	-.9	middle
3	31980	28905	48%	100%	2.72	2.71	1.0	1.60	.01	(2.71)	1.69	1.60	1.62	100%	1.5	highest
										(Mean)		(Modal)	(Median)			

Figure 3. Category Statistics

It appears that collapsing categories 1 and 2, and 4 and 5 optimized the use of the rating scale best because it improved separation and reliability measures and provided better data to model fit than the five category scale (Figure 2). Collapsing middle three categories as an alternative solution did not work. The statistical indicators also showed that combining middle three categories provides poor statistics. Since it is evident that collapsing categories 1 and 2, and 4 and 5 provided the most meaningful information, the data were analyzed on the trichotomous scale.

Facets in the Rasch Model

The extent of the facet contribution to the instructors' score variance was examined. The results are presented for rater facet, instructor facet and item facet respectively as follows.

Rater severity and fit

Figure 4 is a visual representation of Facet analysis results. The first column is the common logit scale, the next three columns present the measures for raters (students) measures and the last column is the scale used in the rating. The second column allows the severity of the raters (students). Their distribution of measures ranged from -7.55 to 9.10 logits severity with a mean of 2.02 and a standard deviation of 2.01. Standard error is .48 with a SD of .47. The results show that majority of logits measures are above zero. This means that majority of students rate their instructor leniently. Moreover, most raters clustered closely around the mean, they are within ±1 logit value. Although the interquartile range is relatively restricted and variability is small, the separation index and reliability was high; 3.93 and .94 respectively. Moreover, the χ^2 of 21348.3 ($p<.000$) was statistically significant and, therefore, the null hypothesis that all raters have the same severity logit estimates must be rejected. In contrast to classical concept of reliability and separation, we do not want high separation index or reliability because we want raters to be equally severe. The separation ratio, 3.93, is an indicator of unwanted variance or construct irrelevant variance and shows that it is 3.93 times greater than the estimation error.

Table 1. Statistics

Statistics	*Examinees*[a] *(Instructor)*	*Raters (Students)*	*Items*
M (measure)	.00[b]	2.02	.00[b]
SD (measure)	1.44	2.01	.25
M (SE)	.17	.48	.03
RMSE	.19	.37	.03
Separation (strata) index (H)	7.69	3.93	7.97
Separation reliability (R)	.98	.94	.98

[a]Examinees with non-extreme scores only
[b]The mean of the measures is constrained in a given facet to be zero.

Outfit and InFit MnSq statistics indicate around 10% of raters had misfit (Table 1). This means that these students rate their instructor inconsistently or consistent with their peers who rate the same instructor. Similarly, around 10% of raters have fit statistics indicating that their ratings are too predictable or provide redundant information. Around 50% of students (611) have InFit value lower than 1.00.

Table 2. Fit Statistics

	Raters (Students)				Instructors				Items			
MnSq	InFit		OutFit		InFit		OutFit		InFit		OutFit	
>1.50	127	(10%)	158	(12.8%)	11	(4.3%)	29	(11.3%)	---		1	(6.7%)
1.5-0.5	980	(80%)	922	(74.7%)	241	(94.6%)	219	(85.9%)	15	(100%)	14	(93.3%)
0.50<	128	(10%)	155	(12.5%)	3	(1.1%)	7	(2.7%)	---		---	

Instructors and fit

The second column in Figure 4 shows teaching effectiveness measure variation among instructors. The instructors are ordered with the most effective at the top and the least effective at the bottom. Measures ranged from -3.58 to 6.45 logits and its distribution is fairly normal around mean of 0.0 with a standard deviation of 1.44; the mean of standard errors of the measures was .17 with a standard deviation of .08. Although the differences in severity are small, the reliability of separation index (7.69) was very high. Instructor separation value is 7.69 that mean this population is separable into 7-8 levels of effectiveness and shows that central tendency effect (Myford & Wolfe, 2004) was not an issue. High separation value provides us high person reliability which is .98. In fact, this coefficient could be a little bit overestimated. For instructor facet, overfit is more of a concern for reliability measure than person estimates. "Overfit tends to stretch the measures along the latent dimension, to reduce their standard errors, and thus, to increase their reliability (or precision); yet, these measures will still be sufficiently accurate for most practical purposes." (p. 102, Eckes, 2015). Fortunately, only small percent (1,1 and 2,7%) of MnSq. values are overfitted.

Items

Table 3 shows statistics and estimates of 15 items in the scale. Observed average of each item is given in the second column of the table on a three category rating scale. It ranged from 2.32 to 2.5. While Item 12 is the easiest to endorse, Item 15 are the hardest to endorse item. Fair average (column 3) is a transformed score of Rasch measures (http://www.winsteps.com/facetman/fairaverage.htm). The items were set to have mean of zero logit. Difference in item measures does not reflect a substantial difference in difficulty. The range is from -.40 to .39 with a mean of .00 and a standard deviation of .25. The fifth column shows the amount of error corresponding to each measure. The error was equal to .03 for all items. Item fit indexes, Infit and Outfit MnSq values are all within acceptable range except Item 10. In fact, Outfit MnSq value is just above acceptable range,

1.77 and Infit MnSq value is in the acceptable range. Linacre argues that it is easy to find misfit when the data size is big enough (http://www.winsteps.com/winman/globalfitstatistics.htm). As a result, we can consider Item 10 has acceptable fit as well.

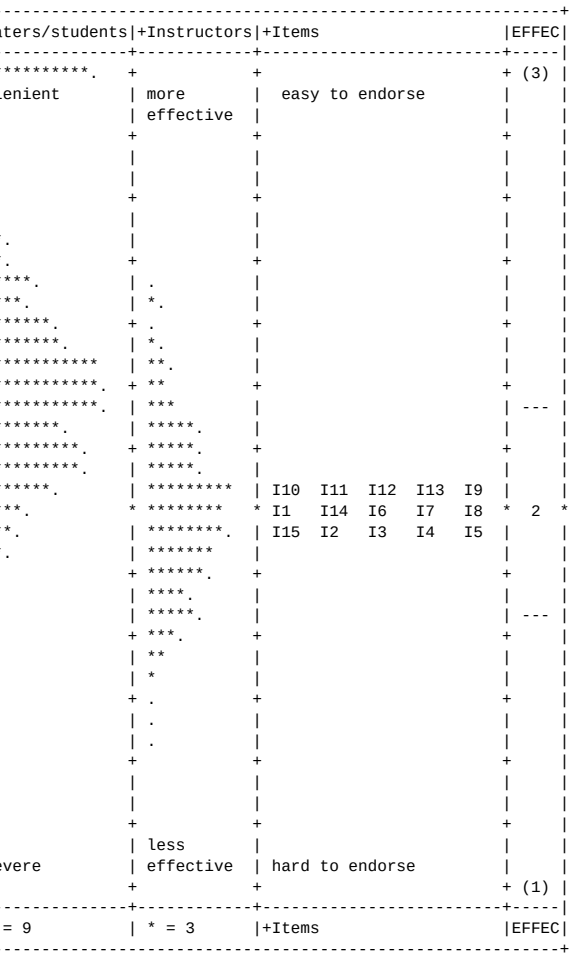


Figure 4. Facets Summary (Rater (students) Severity, Instructor Effectiveness, Item Difficulty)

Table 3. Item Statistics and Estimates

ITEM	Observed average	Fair average	Measure (in logits)	SE	InFit MnSq	OutFit MnSq
15	2.32	2.41	-0.40	0.03	1.31	1.31
5	2.34	2.43	-0.34	0.03	0.87	0.95
2	2.35	2.44	-0.30	0.03	0.91	1.06
4	2.35	2.44	-0.3	0.03	0.9	0.93
3	2.37	2.47	-0.19	0.03	1.05	1.16
14	2.4	2.50	-0.07	0.03	0.86	0.84
7	2.41	2.52	-0.01	0.03	0.98	0.93
1	2.42	2.53	0.01	0.03	0.93	1
8	2.42	2.53	0.02	0.03	1.05	1.06
6	2.45	2.57	0.14	0.03	0.88	0.87
9	2.46	2.58	0.19	0.03	1.04	1.25
10	2.46	2.58	0.20	0.03	1.2	1.77
13	2.48	2.61	0.32	0.03	1	0.98
11	2.49	2.62	0.34	0.03	1.02	1.08
12	2.5	2.63	0.39	0.03	0.98	1.44

Bias Analyses

The second purpose of the study is bias analysis. The students' judging behavior is examined by using the Many-Facet Rasch Model in order to detect any potential source of bias in student evaluation of teaching.

MFRM uses the term 'bias' differently from its meaning in the measurement literature. A bias analysis (an interaction analysis) helps to pinpoint unexpected response pattern by considering more than one facet at the same time. If there is a deviation from what was expected, these patterns point the bias (interactions). Two two-way interaction analyses were conducted; item by course type and item by expected grade.

The course type facet with two elements (required and elective course) was added to my basic three facet model. Bias diagram of course type bias illustrated in Figure 7 in the appendix. As it is seen that students in elective courses rated instructors more positively than the students in required courses. However, a logit measure difference between two course types is not substantial; it is -.07 and .07 for must course and elective course respectively. Interaction analysis indicates that students are able to keep their severity consistently across items on each course type and no evidence of bias were observed in any of the 30 combinations. In other words, instructors of elective courses got always higher ratings on each item than the instructors of the must course did. Figure 8 in the appendix shows bias diagram of t-values which are all within ±2 [χ^2 (30)=7.5, p=1.0].

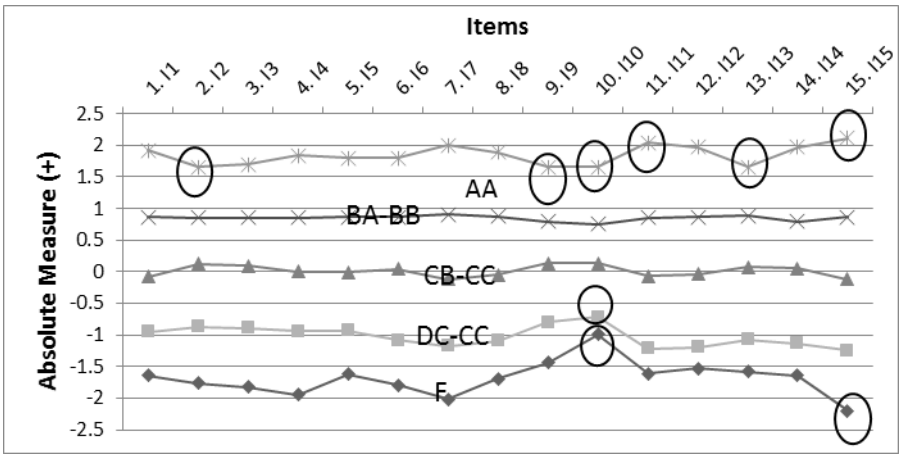


Figure 5. Bias Diagram: Between Items and Expected Grade

Expected grade facet with five elements (AA, BA-BB, CB-CC, DC-DD and F) was added to my basic three facet model. The highest and lowest element difference between measures was 3.52 logits. As the expected grade gets higher, the average rating of the instructors gets higher. Separation reliability of 1.00 shows that this average rating of measures significantly differs across elements of

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq	N	Expected Grade Items	measr	Nu	Ite	measr
183	166.92	89	.18	.69	.21	3.34	88	.0012	1.3	1.3	46	1	F	-1.68	10	I10	-.22
604	582.16	287	.08	.30	.12	2.54	286	.0118	1.1	1.1	47	2	DD DC	-1.02	10	I10	-.22
2538	2489.97	964	.05	.27	.08	3.54	963	.0004	1.3	1.5	75	5	AA	1.84	15	I15	.43
2649	2620.68	964	.03	.20	.09	2.35	963	.0189	1.0	1.5	55	5	AA	1.84	11	I11	-.36
2568	2596.28	964	-.03	-.18	.08	-2.28	963	.0227	1.1	1.1	45	5	AA	1.84	9	I9	-.20
2590	2617.19	964	-.03	-.18	.08	-2.25	963	.0249	1.1	1.1	65	5	AA	1.84	13	I13	-.34
2476	2509.99	964	-.04	-.19	.07	-2.54	963	.0113	1.0	1.9	10	5	AA	1.84	2	I2	.32
2570	2599.14	964	-.03	-.19	.08	-2.36	963	.0185	1.3	2.9	50	5	AA	1.84	10	I10	-.22
141	152.20	89	-.13	-.52	.22	-2.37	88	.0199	1.8	2.2	71	1	F	-1.68	15	I15	.43

Figure 6. Bias/Interaction, Facet Output

the expected grade facet. Expected grade contributes the variance in the model as much as instructors' ability does. Almost 40% of the variance could be explained by the students' expected grade. Bias diagram of expected grade illustrated in Figure 6 shows that there is an interaction between items and expected grades. Nine combination of facet elements out of 75 have a z-score equal or greater than 2. Those elements with statistically significant interactions were shown within a circle in Figure 5 [χ^2 (75)=139.3, p=.00].

Figure 6 provides statistics for nine significant interactions. These elements are shorted according to bias size which is a maximum value of .69 and a minimum value of .18. Students who expected AA from the course rated Item 11 and 15 unexpectedly higher than the models expected. On the contrary, they rated Item 2, 9, 10 and 13 unexpectedly lower. Likewise, students who expected grade lower than CC overrated Item 10 and students expecting to fail the class underrated item 15.

DISCUSSION and CONCLUSION

Economic and political changes in the World have been pushing higher education institutes to exhibit their performances equally well in not only research and but also teaching. European Association for Quality Assurance in Higher Education (2009) emphases that higher education institute should monitor qualification and competence of teaching staff. Although it is widely accepted that SET should not be sole toll to evaluate ones teaching quality, SET result will continue to be used for longer time as internal quality assurance of teaching effectiveness (Penny, 2003) in spite of all arguments against it.

Although student evaluation of teaching has been implementing in western higher education institutes since nearly the beginning of the 20th century, few universities in Turkey have had adopted this evaluation system. This study used the data set obtained from one of these few universities implementing SET. The purpose of this study is to examine what extent do the facets (instructor, student, and rating items) modeled in instructor evaluation contribute to instructors' score variance and examine the students' judging behavior using the MFRM to examine any potential source of bias in student evaluation of teaching effectiveness.

MFRM provides a stronger measurement model than any other method in evaluation of rater mediated assessment. MFRM offers several statistics that helps us examine what extent instructor, student, and rating items in instructor evaluation contribute to instructors' score variance and examine the students' judging behavior using the MFRM to examine any potential source of bias in student evaluation. Rater effect in an evaluation process appears in different forms such as such as severity or leniency, halo or central tendency (Hoyt, 2000; Myford & Wolfe, 2003). Such rater effect introduces a method variance in observed ratings which are associated with the raters and not with examinee. MFRM provides us statistics to evaluate the extent of rater effects. Some of statistics utilized in this study are reliability and separation index, logit measures and Infit and Outfit MnSq. They are discussed respectively in this section.

The interpretation of these statistics depends on the facet considered. Given a small range of logit measures (± 2 logits) separation reliability and separation index for instructor facet is surprisingly high, .95 and 7.69. This result indicates that the spread of the effectiveness measures was considerably much greater than the precision of those measures and most probably big sample sizes resulted high separation among instructors. In general performance assessment it is aimed to differentiate among examinees, therefore, high separation reliability is desired.

The separation reliability and separation index are .92 and 3.93 for student (rater) facet. Unlike instructors (examinee) facet, we do not want high statistics for this facet because ideally we wish equal leniency or severity for raters. When raters practice a highly similar degree, reliability becomes low. Therefore, for this facet low reliability and separation is desired. These statistics showed students differed strongly in the severity with which they rated instructors. In overall, students display a strong leniency effect. This result is supported by Zhao and Gallant (2012).

For item facet, the range of difficulty measures is too small (± .5 logits) and approximately 50% of students answer to items are too predictable, this means that students rated each criteria (item) quite similarly. This result can be interpreted in different ways. These 15 items in the questionnaire is redundant, they almost provide the same information about instructors. Therefore, some items can be eliminated and extra items could be added to widen the separation index among items. A halo effect, another source of rater effect, is signaled by these group level and individual level statistics. "A halo effect refers to a rater's tendency to provide similar ratings of an examinee's performance on conceptually distinct criteria…When the majority of the raters were subject to halo error, the ratings would be highly similar across criteria and, as a result, the criteria showed only little variation in their measures of difficulty. " (Eckes, 2011, p.66). Myford and Wolfe (2004) indicated that rater Infit and Outfit MnSq indices less than one or greater than one, depending on measurement context can be used to diagnose a halo effect. Approximately 50 percent of rater has Infit and Outfit MnSq, values less than one. Low variability in item difficulty measure and large number of MnSq values lower than one draw attention to possible a halo effect. It appears that ratings are highly redundant across criteria.

Another possible rater effect is a central tendency effect. It happens when raters tent to overuse the middle categories of the rating scale. In case of central tendency effect, the scale is only functional for average performing examinees, not with low performing or high performing examinees (Eckes, 2011). This kind of rater effect is not an issue in this study. However, the five point rating scale does not work as expected. Students are clustered at the high end of the five point scale. After rescaling, the results are still similar. Therefore, as a group, students display a strong leniency effect.

So far each facet of the Many Facet Rasch Model was singled out and discussed. The last research question is about potential biasing of the SET questionnaire with respect to course type and expected grade. Exploratory two way interaction analysis, it is also known as bias analysis was used to identify systematic deviations from expectations. The interaction between course type and item facet is not statistically significant. Students in elective courses rated instructors more positively than the students in required courses. Instructors who teach elective courses always get higher average score then instructors teaching required courses. In the second bias analysis, the interaction between students' expected grade facet and item facet was examined. As the expected score gets higher, the instructor score gets higher. Moreover, there is interaction between them. The students expecting AA give the highest score to Item 15 "Overall effectiveness of the instructor" and even higher than the expected score by the model. In contrast, students expecting to fail a course give the lowest average score and even lower than the expected. The finding related to expected grade is supported by previous research (e.g. Dodeen, 2013; Marks, 2000; Marsch, 2007).

In conclusion: Many researchers (e.g. Dodeen (2013), Gursoy and Umbreit (2005), Marks (2000), Marsch, 2007) states that effective teaching is a multidimensional construct. On the other hand, the SET items of this study display unidimensional psychometric structure. Spooren, Brock and Mortelmans (2013) concluded that use of SET in higher education and validity of the scores obtained with SET should continue to be questioned. My conclusion is similar to them. It looks like the most serious threat in SET is halo effect. Halo effect shows that students do not evaluate their instructor as we expected. While evaluating their instructors, they may have different criteria in their minds other than the criteria that the university sets for their instructors.

REFERENCES

Abrami, P. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research, 2001*(109), 59-87. http://dx.doi.org/10.1002/ir.4

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 357-74.

Barnes, D. C., Engelland, B. T., Matherine, C. F., Martin, W. C., Orgeron, C. P., Ring, J. K., et al. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, *42*(1), 199-213.

Basow, S. A., & Martin, J. L. (2012). Bias in student ratings. In M.E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators*. Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/evals2012/index.php

Beran, T., Violato, C., & Kline, D. (2007). What's the 'use' of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education, 17*(1), 27-43.

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited. IDEA Paper No. 32.* Retrieved from http://www.faculty.umb.edu/pjt/cashin95.pdf

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass.

Dodeen, H. (2013). Validity, reliability, and potential bias of short forms of students' evaluation of teaching: The case of UAE University. *Educational Assessment, 18*, 235-250.

Eckes, T. (2005). Examining rater effects in Testdaf writing and speaking performance assessments: A Many-Facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Eckes, T. (2009). Many-Facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt am Main: Peter Lang.

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603–615.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models, and trends.* Toronto: Higher Education Quality Council of Ontario.

Gursoy, D., & Umbreit, W. T. (2005). Exploring students' evaluations of teaching effectiveness: What factors are important? *Journal of Hospitality & Tourism Research, 29*, 91-109.

Haladyna, T., & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education, 35*, 1209–1217.

Hoyt, D. P., Chen, Y., Pallett, W. H., & Gross, A. B. (1999). IDEA Technical Report No. 11: Revising the IDEA systems for obtaining student ratings of instructor and courses. Kansas State University, Manhattan, KS. The IDEA Center. Retrieved from http://ideaedu.org/wp-content/uploads/ 2014/ 11/techreport-11.pdf.

Hoyt, W. (2000). Rater bias in psychological research: When is it a problem and what can we do about it?. *Psychological Methods, 5*(1), 64-86.

Hoyt, W., & Kerns, M. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403-424.

Keeley, J. (2012). Course and instructor evaluation. In W. Buskist & V. A. Benassi (Eds.), *Effective college and university teaching. Strategies and tactics for the new professoriate* (pp. 173-180). Thousand Oaks, CA: Sage.

Koh, H., & Tan, T. (1997). Empirical investigation of factor affecting SET results. *International Journal of Educational Management, 11*, 170-208.

Lane, S., & Stone, C. A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational Measurement*. Westport, CT: American Council on Education & Praeger

Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago: MESA.

Linacre, J. M. (1994). Constructing measurement with a many-facet Rasch model. In M. Wilson (Ed.) *Objective measurement: Theory in practice* (Vol. 2, pp. 129-144) Norwood, NJ: Abex.

Linacre, J. M. (2009). FACETS (Computer program, version 3.66.1). Chicago: MESA.

Linacre, J. M., & Wright, B. D. (2002). Understanding Rasch measurement: Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486-512.

Lunz, M., Wright, B., & Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345.

Marks, R. (2000). Determinants of student evaluations of global measures of instructor and course value. *Journal of Marketing Education, 22*(2), 108-119.

Marsh, H. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*(1), 77-95.

Marsh, H. W. (1984). Students' evaluation of university teaching: Dimensionality, reliability, validity, potential bias, and utility. *Journal of Educational Psychology, 76*, 707–754.

Marsh, H. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-388.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187–1197.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35-44.

Moore, S., & Kuol, N. (2005). Students evaluating teachers: Exploring the importance of faculty reaction to feedback on teaching. *Teaching in Higher Education, 10*(1), 57-73.

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies, 35*, 547–52.

Mulqueen, C., Baker, D. P., & Key Dismukes, R. (2002). Pilot instructor training: The utility of the multifacet Item Response Theory model. *International Journal of Aviation Psychology, 12*, 287-303.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Nelson, J. P., & Lynch, K. A. (1984). Grade inflation, real income, simultaneity, and teaching evaluation. *Journal of Economic Education, 15*, 21–39.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Teaching and Learning, 5*, 27-44.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.

Penny, A. R. (2003) Changing the agenda for research into students' views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*(3), 399-411.

Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education, 39*(46), A40.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598–642.

Sudweeks, R., Reeve, S., & Bradshaw, W. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*(3), 239-261.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-211.

Williams, W. M., & Ceci, S. (1997). "How'm i doing?" Problems with student ratings of instructors and courses. *Change: The Magazine of Higher Learning, 29*(5), 12-23.

Wright, R. (1996). A study of the acquisition of verbs of motion by Grade 4/5 early French immersion students. *The Canadian Modern Language Review, 53*(1), 257-280.

Wright, B., & Stone, M. (1979). *Best test design*. Chicago: Mesa Press.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching In Higher Education, 12*(1), 55-76.

Zangenehzadeh, H. (1988). Grade inflation: A way out. *Journal of Economic Education, 19*, 217–230.

Zhao, J. Z., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227-235.

Acknowledgements

This work was supported by B.U. Research Fund at Boğaziçi University under contract 14D05P1

UZUN ÖZET

Giriş

Ders değerlendirme anketi (DDA), farklı amaçlar için akademik dönemin herhangi bir noktasında öğrencilerin ders ve öğretim elemanı hakkındaki deneyimlerine dair görüşlerini toplamak amacıyla sıkça kullanılmaktadır (Penny, 2003, Seldin, 1993; Zabaleta, 2007). DDA sonuçları yardımıyla, yüksek öğretim kurumlarında yöneticiler (a) öğretim niteliğini artırmayı, (b) yükseltme veya ödül gibi karar noktalarında veri sağlamayı ve (c) kurumsal hesap verebilmek amaçlı kanıt sağlamayı hedeflemektedirler (Seldin, 1993 ; Spooren, Brock ve Mortelmans, 2013).

DDA yardımıyla yapılan değerlendirmeler bir tür performans değerlendirme olarak kabul edilebilir. Performans değerlendirmede, değerlendirmeye tabi olan kişi bir performans gerçekleştirir ve/veya bir ürün oluşturur ve bu performans veya ürünün kalitesi, en az bir değerlendirici tarafından puanlanır. Performans değerlendirme, bir puanlayıcı aracılı değerlendirme süreci olduğundaysa ekstra önlemler daha güvenilir ve adil bir değerlendirme uygulamaları oluşturmak için dikkate

alınması gerekir. Değerlendirici aracılı değerlendirmede en yaygın tehditlerden birisi 'değerlendirici varyansı"dır. Bu terim, değerlendirilen kişinin performansının kendi beceri seviyesinin yanı sıra değerlendiricinin özelliklerine (katılık/cömertlik, cinsiyet gibi) bağlı olmasına karşılık gelir (Eckes, 2009). Başka bir değişle, ölçmek istediğimiz yapıyla ilgisiz varyanstan oluşan değerlendirici varyansı, performans değerlendirmenin adaletini ve geçerliğini tehdit eder (Messick, 1998; Lane & Stone, 2006).

Alan yazımının gösterdiği gibi, DDA'ya ilişkin teorik ve psikometrik tartışmalar süregelmektedir (Gravestock & Gregor-Greenleaf, 2008). Akademik çalışmalar aslında birbiriyle ilişkili iki ana kaygı etrafında toplanmıştır. Birincisi kaygı elde dilen puanların geçerliğiyle ilgilidir. DDA alan yazınımda "Etkili öğretim" diye tanımlayabileğimiz yapıyı ne derece ölçtümüz soru işaretidir. İkinci kaygıysa elde edilen puanlarda ortaya cıkabilecek ve puanların geçerliğini ve güvenirliğini tehdit edebilecek yanlılık kaynaklarıdır (Gürsoy & Umbreit, 2005). Bütün bu bağlam içinde, bu çalışmanın amacı, a) öğretim elemanlarının puanlarındaki farklılığa/varyansa, değerlendirme sürecindeki elemanların (öğretim elemanı, öğrenci ve değerlendirme maddeleri) ne derece katkı sağladığını ve b) öğrencilerin değerlendirmelerinde yanlılığa yol açacak potansiyel kaynakları çok yüzeyli Rasch modeli yardımıyla incelemektir.

DDA'in, 1920 yılında Washington Üniversitesi'nde ilk kullanımından bu yana oldukça büyük ölçüde alan yazımı oluşmuştur. O zamandan bu yana, DDA ile elde edilen puanların geçerliliği tartışma konusu olduğu gibi, DDA kullanımının eğitimin niteliğini artırıp artırmadığı gibi konular yeni tartışma konuları olarak alana girmiştir. Akademik araştırmaların çoğunluğu Kuzey Amerika, Avustralya ve İngiltereki yüksek öğrenim bağlamı içinde yapılmıştır (Gravestock & Gregor-Greenleaf, 2008; Zabaleta, 2007). Bu çalışmaların çoğu (Abrami, 2001; Beran, Violato & Kline, 2007; Gravestock, & Gregor-Greenleaf, 2008; Marsh, 1987 gibi) DDA'nin kullanımları için genel olarak olumlu bir tutuma sahiptir; Öte yandan, bazı araştırmacılarda nitelikli öğretimle ilgisi olmayan ders ve öğretim elemanı özellikleri yüzünden yanlı sonuçlar verebileceğinden dolayı DDA'nın kullanımında şüpheci bir tutum sergilemektedirler (Dedeen, 2013; Koh & Tan, 1997; Williams & Ceci, 1997). Görüldüğü gibi alan taraması birbirleriyle çelişecek sonuçlar vermektedir. Bu nedenle, DDA sonuçları çok önemli kararlarda, işe alma, promosyon veya yükseltmelerde tek başına kullanılmaması gerektiği düşünülmektedir.

Bir performans değerlendirmede, kişinin puanı bu süreçteki bir grup aktöre bağlıdır. Bunlardan en sık görülenleri; performans görevini alan kişinin beceri seviyesi, performans görevinin zorluk derecesi (Mulqueen, Baker & Dismukes, 2002) ve puanlayıcı etkisidir. Değerlendirme sürecindeki puanlayıcı etkisi farklı sekillerde ortaya çıkabilir. Bunlardan bazıları, katılık/cömertlik, halo etkisi, ve merkezi eğilim etkisidir (Hoyt, 2000; MyFord & Wolfe, 2003). Bu tür puanlayıcı etkileri gözlenen puanlarda metot varyansı oluşturur ve bu varyans performans görevini alan kişiyle ilgili değil, puanlayıcı ile ilgilidir. Farklı bağlamlarda yapılan arastırmalar gösteriyorki puanlayıcı aracılı performans değerlendirmelerinde puanlayıcı etkisi çok fazladır (Eckes, 2005). Örneğin, Hoyt ve Kerns (1999) meta analiz araştırmasında performans görevisi alanların performanlarının %37'si puanlayıcı etkisi ve puanlayıcı-sınavı alan kişi arsındaki etkileşimle açıklanabilir. Birden fazla puanlayıcının olduğu değerlendirme süreçlerinin güvenirliğini test etmek için daha standart prosedürlerden, modern test teorilerinin faydalanıldığı bir dizi yöntem vardır.

Yöntem

Bu çalışmada, büyük bir şehirinde bulunan orta boy bir devlet üniversitenin lisans derslerinden ders değerlendirme anketleriyle toplanan ders ve öğretim elemanı değerlendirme verilerini kullanılmıştır. DDA üç bölümden oluşmaktadır: İlk bölüm ders ve tasarımı hakkında beş madde ve öğretim elemanın etkinliği hakkında ise 10 madde içerir. Her madde beş puanlaı derecelendirme ölçeğine (1, Mükemmel: 5, Kötü) sahiptir. İkinci bölümde ise derse katılım, dersten beklenen not ve dersin programdaki türü (zorunlu/seçmeli) gibi bilgileri ölçen maddeler yer almaktadır. Son bölümde ise öğrencilerin ders ve öğretim elemanıyla ilgili geri bildirimlerini yazabilmeleri için bir metin kutusu sağlanmıştır.

Veri setinde bağlantıyı kurabilmek için tüm üniversiteye çok sayıda ders açan Fen Edebiyat Fakültesi amaçlı olarak seçilmiştir. 2015 Güz Akademik döneminde, DDA uygulandığı ve en az altı öğrencinin katıldığı 254 dersten 1.235 öğrencinin verisi analiz edilmiştir. Rasch analizinden önce, tek boyutluluk varsayımı doğrulayıc ve açımlayıcı faktör analiziyle incelenip, doğrulanmıştır. Rasch analizleri Facet v. 3.71.4 (Linacre , 1987-2014) kullanılarak tamamlanmıştır. Rasch Derecelendirme Ölçeği Modeline (1978 Andrich) dayanan bir üç yüzeyli, iki adet dört yüzeyli modeller kullanılmıştır.

Sonuç ve Tartışma

Analizde kullanılan toplam 63.811 verinin standardlaştırılmış değerinin ±3'den büyük ya da eşit olanlarının sayısı 614 (%0.96), ±2'den büyük olanların sayısı ise 2.870 (%4.49) olarak elde edilmiş ve model veri uyumu sağlanmıştır.

Puanlayıcı yüzeyi incelendiğinde; öğrencilerin farklı katılık derecesine sahip oldukları görülmektedir. Ayırma güvenirliği ve indeksi 0,92 ve 3,93'dür. İdeal durumda puanlıyıcıların eşit katılık derecesine sahip olması beklenir.

Anket maddeleri incelendiğinde maddelerin zorluk dereceleri ±0,5 logit arasında değiştiği ve öğrencilerin yaklaşık %50'sinin cevapları oldukça tahmin edilebilir olduğu görülmüştür. Bu durum iki ayrı şekilde yorumlanabilir. Bu maddeler neredeyse aynı bilgiyi sağlamaktadır bu nedenle bu 15 maddenin hepsi gerekli değildir. Bazı maddeler çıkarılırken, ayırma indeksini yükseltecek şekilde yeni maddeler eklenebilir. Bunun yanı sıra, madde zorluk derecelerindeki düşük varyazyon, grup seviyesindeki muhtemel alo etkisinide de dikkat çekmektedir. Myford and Wolfe (2004) uyguluk içi ve dışı istatistik değerlerinin ölçmede birey seviyesinde halo etkisini belirlemek için kullanılabileceğini belirtmişlerdir. Madde zorluk değerlenlerindeki düşük varyans ve yüksek sayıda birden küçük öğretim elemanı uyum indeks değerleri muhtemel halo etkisine dikkat çekmektedir.

Diğer muhtemel puanlayıcı etkisi ise merkezi eğilim etkisidir. Merkezi eğilim etkisi, puanlayıcı ölçeğin orta kategorilerinin gerektiğinden fazla kullanılmasıyla ortaya çıkar. Bu tür etki gözlenmemektedir. Fakat beş puanlı ölçek istenilen şekilde çalışmamaktadır. Öğrencilerin ölçeğin üst kategori çok kullandığı görülmektedir. Bu durumda öğrencilerin bol notlu davranışa sahip olduklarını göstermiştir.

Farklı yüzeyler arasındaki etkileşime bakıldığında, maddelerle ders tipi (seçmeli/ zorunlu) arasında bir etkileşim olmadığı belirlenmiştir. Fakat maddelerle mğrencinin beklediği not arasında bir etkileşim bulunmuştur. Dersten AA bekleyen bir öğrenci Madde 15'e beklenenden yüksek puan verirken, dersten kalmayı bekleyen öğrenci beklenenden daha düşük puanlama yapmıştır.

APPENDIX

A Three Facet Model

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \tau_k \quad \text{where}$$

P_{njimk}, = probability of instructor *n* receiving a rating of *k* on criterion *i* from student *j*,
$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,
β_n = ability of person *n*,
δ_i = difficulty of criterion (item) *i*,
γ_j = severity of rater *j*,
τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

The first four Facet model

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \pi_m - \tau_k \quad \text{where}$$

P_{njimk}, = probability of instructor *n* receiving a rating of *k* on criterion *i* from student *j*,
$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,
β_n = ability of person *n*,
δ_i = difficulty of criterion (item) *i*,
γ_j = severity of rater *j*,
π_m = severerity of course type m
τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

The second four Facet model

$$\ln\left(\frac{P_{nij}}{P_{ni(k-1)}}\right) = \beta_n - \delta_i - \gamma_j - \pi_m - \tau_k \quad \text{where}$$

P_{njimk}, = probability of instructor *n* receiving a rating of *k* on criterion *i* from student *j*,
$P_{njim(k-1)}$ = probability of person *n* receiving a rating of *k*-1 on criterion i from rater *j*,
β_n = ability of person *n*,
δ_i = difficulty of criterion (item) *i*,
γ_j = severity of rater *j*,
π_m = severity of expected grade m
τ_k = difficulty of receiving a rating of *k* relative to a rating of *k*-1.

SET Questionnaire: The content of the items

1. Course objectives
2. Course design
3. Course materials
4. Course requirements and assignments
5. Overall effectiveness of the course
6. Course materials
7. Awareness of students' comprehension
8. Encouragement of student participation in class
6. Effective use of class time
7. Grading practices
8. Fair grading
9. Fair handling of objections to grades
10. Availability to help
11. Overall effectiveness of the instructor
12. I would choose to take another course with the same instructor

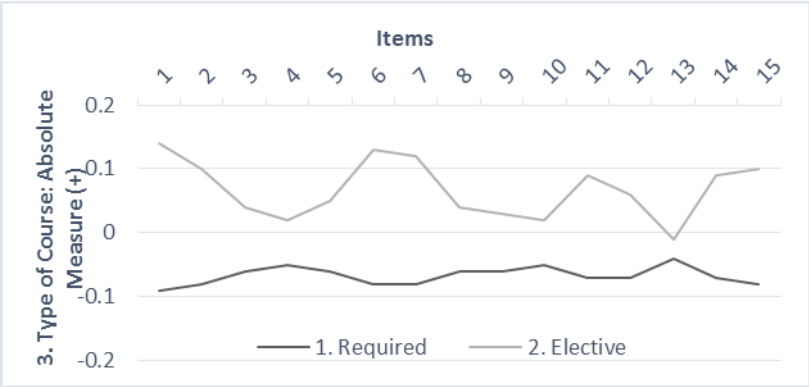


Figure 7. Bias Diagram Showing the Interaction Between Items and Course Type

Figure 8. Bias Diagram Showing the Interaction Between Items and Course Type
Series 1=required 2=elective

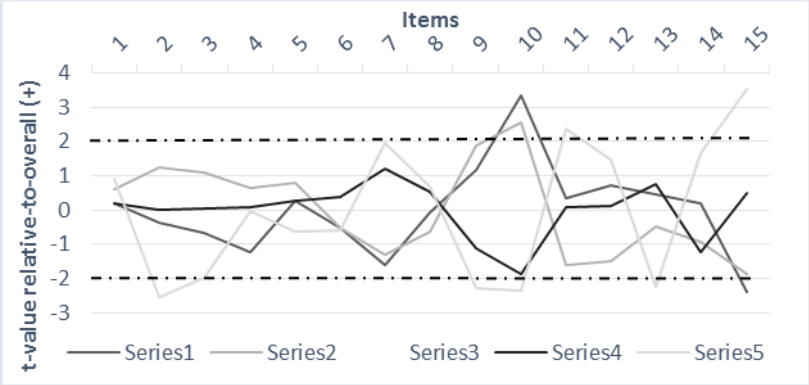


Figure 9: Bias diagram showing the interaction between items and expected grade
Series 1=F, 2=DC-CC, 3=CB-CC, 4=BB-BA, 5=A

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);34-46

Basit ve Karmaşık Test Desenlerinde Çok Boyutlu Madde Seçme Yöntemlerinin Karşılaştırılması*

A Comparison of Multidimensional Item Selection Methods in Simple and Complex Test Designs

Eren Halil ÖZBERK ** Selahattin GELBAL ***

Öz

Bu araştırmada diğer araştırmaların aksine toplam yetenek puanları gerçek test koşullarına uygun olacak şekilde farklı test koşullarında karşılaştırılmıştır (basit ve karmaşık). Araştırmada test deseni, boyut başına düşen soru sayısı, boyutlar arası korelasyon ve madde seçme yöntemleri olmak üzere dört koşul manipüle edilmiştir. Veri setleri, üretilen madde ve yetenek parametreleri ve M3PL telafi edici çok boyutlu madde tepki kuramı modeli kullanılarak belirlenen korelasyonlara bağlı kalarak üretilmiştir. Çok boyutlu bireyselleştirilmiş bilgisayarlı test uygulamaları sonucu elde edilen toplam yetenek puanları mutlak yanlılık (ABSBIAS), korelasyon ve hata kareleri ortalamasının karekökü (RMSE) kullanılarak karşılaştırılmıştır. Sonuçlar incelendiğinde çok boyutlu test deseni, boyut başına düşen madde sayısı ve boyutlar arası korelasyon değişkenlerinin toplam puanları kestirmede madde seçme yöntemleri üzerinde etkilerinin olduğu belirlenmiştir. Basit yapıdaki bir test için Minimum Hata Varyansı madde seçme yönteminin hem uzun hem de kısa testler için en düşük mutlak yanlılık değerinin ürettiği belirlenmiştir. Model karmaşıklaştıkça Kullback-Leibler madde seçme yönteminin diğer iki yöntemden daha iyi performans gösterdiği belirlenmiştir.

Anahtar Kelimeler: Madde seçme yöntemi, çok boyutlu bireyselleştirilmiş bilgisayarlı test, çok boyutlu maddde tepki kuramı, toplam puan kestirimi

Abstract

In contrast with the previous studies, this study employed various test designs (simple and complex) which allow the evaluation of the overall ability score estimations across multiple real test conditions. In this study, four factors were manipulated, namely the test design, number of items per dimension, correlation between dimensions and item selection methods. Using the generated item and ability parameters, dichotomous item responses were generated in by using M3PL compensatory multidimensional IRT model with specified correlations. MCAT composite ability score accuracy was evaluated using absolute bias (ABSBIAS), correlation and the root mean square error (RMSE) between true and estimated ability scores. The results suggest that the multidimensional test structure, number of item per dimension and correlation between dimensions had significant effect on item selection methods for the overall score estimations. For simple structure test design it was found that V1 item selection has the lowest absolute bias estimations for both long and short tests while estimating overall scores. As the model gets complex KL item selection method performed better than other two item selection method.

Keywords: Item selection method, multidimensional computer adaptive testing, multidimensional item response theory, composite score estimation

GİRİŞ

Eğitim ve psikoloji alanında değerlendirme araçlarının başlıca amacı ölçülen özeliğin miktarını belirlemek ve elde edilen numerik puanları kullanarak bireyleri örtük özelliklerine göre sıralamaktır. Puanlar, sıralama amacıyla kullanıldığı durumlarda önemli bir değerlendirme ölçütü olarak olabilmektedir. Özellikle başarı düzeylerinin belirlenmesinde, sertifika ve lisanslama yetkilerinin

* Bu çalışma, birinci yazarın Prof. Dr. Selahattin GELBAL danışmanlığında tamamlanan doktora tezinden türetilmiştir.
** Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, eposta: erenozberk@gmail.com
*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, eposta: sgelbal@gmail.com

DOI: 10.21031/epod.286956
Geliş Tarihi: 22.01.2017
Kabul Tarihi: 06.03.2017

verilmesinde puanlar önemli bir sıralama ölçütüdür. Örneğin bir öğrencinin üniversiteye giriş sınavından elde ettiği puan öğrencinin bilgi ve becerisinin bir göstergesidir.

Toplam yetenek puanlarının rapor edildiği durumlarda, tek boyutlu test desenleri yerine kullanılan çok boyutlu test desenlerinin parametre kestirimlerinde önemli rol oynamaktadır. Çok boyutlu madde tepki kuramı (ÇBMTK) modelleri test yapıları bakımından basit ve karmaşık yapı olarak ikiye ayrılmıştır. Literatürde bunun için birçok farklı adlandırma mevcuttur. Bazı araştırmacılar basit ve karmaşık yapılı test deseni olarak adlandırırken (Luo, 2013; Yao, 2012; Zhang, 2012) bazıları maddeler arası model (multidimensional between-item model) ve maddeler içi model (multidimensional within-item model) olarak adlandırmıştır (Adams, Wilson ve Wang, 1997; Bulut, 2013; Wang, Chen ve Cheng, 2004). Maddelerin faktöriyel karmaşık bir yapıda olması durumunda belirli maddeler birden fazla boyuta yük verebilirler ve yine birden fazla boyuta bilgi sağlayabilirler. Testlerde yer alan maddeler çok boyutlu özelliğe sahip olduğunda çok boyutlu ölçme bilgisinin ortaya çıkması, alt boyut ve toplam yetenek puanlarının kestirimine etki edebilmektedir (Bulut, 2013; Liu, 2015; Luecht, Gierl, Tan ve Huff, 2006; Luo, 2013; Zhang, 2012; Finch, 2010). Yapılan araştırmalarda çok boyutlu test yapılarının çok boyutlu bilgi düzeylerini farklılaştırdığı, bu bakımdan elde edilen yetenek ve madde parametrelerini etkilediği görülmüştür. ÇBMTK modellerinin verdiği avantajlar göz önüne alındığında bazı araştırmacılar gerçek test koşullarına yakın olması nedeniyle karmaşık çok boyutlu yapıların kullanılmasını önerirken (Ackerman, 1994; Reckase, 2009) diğer araştırmacılar basit yapıdaki desenler kullanmanın karmaşık yapıda desenler kullanmadan daha avantajlı olduğunu belirtmişlerdir (Luecht & Miller, 1992; Yao & Boughton, 2009).

Çok boyutlu bireyselleştirilmiş bilgisayarlı testler (ÇBBBT), hem alt boyut hem de toplam puanlar rapor etmede tek boyutlu bireyselleştirilmiş bilgisayarlı test (TBBBT) uygulamalarına göre avantajlı durumdadır. ÇBBBT uygulamalarında her bir alt yeteneğe göre puanlar daha az madde ile kestirilebilmektedir ve her boyuta ait puanlar rapor edilebildiğinden bireylerin zayıf ve güçlü yanları boyutlara göre rahatlıkla belirlenebilmektedir (Wang ve Chang, 2011). Yakın zamanda yapılan araştırmalarda TBBBT uygulamaları için geliştirilen birçok madde seçme yöntemi ÇBBBT uygulamalarına göre tekrar geliştirilmiştir (Mulder ve van der Linden, 2010; Segall, 1996; van der Linden, 1999; Veldkamp ve van der Linden, 2002). TBBBT ve ÇBBBT madde seçme yöntemleri kestirdikleri yetenek sayılarının farklı olması bakımından ayrılmaktadır. Özellikle ÇBBBT madde seçme yöntemleri, çoklu yeteneklerin kestirimlerine getirdikleri farklı teknikler ile alt boyut puanlarının hesaplanmasında daha kararlı sonuçlar elde etmeye çalışmışlardır. Örneğin, Segall (1996) tarafından önerilen madde seçme yöntemi, genel varyansı azaltarak yeteneklere ait güven aralıklarını düşürmek isterken, van der Linden (1999) tarafından önerilen yöntem ise her bir yetenek kestirimine ait toplam varyansı azaltma yoluna gitmiştir.

Çok Boyutlu Bireyselleştirilmiş Bilgisayarlı Testler

Bireyselleştirilmiş testlerde çok boyutlulukla ilgili ilk çalışmalar Bloxom ve Vale (1987), Fan ve Shu (1996), Luecht (1996), Segall (1996) ile başlamış daha sonraları van der Linden (1999; 2005), Mudler ve van der Linden (2009) ile devam etmiştir. Yapılan ilk çalışmalar yetenek kestirimleri ve madde seçme yöntemleri üzerine yoğunlaşmıştır. Daha sonraları Wang ve Chen (2004), çok boyutlu test yapılarının yetenek kestirimleri üzerindeki etkilerini incelemişledir. Yakın zamanda yapılan çalışmalar, madde seçme yöntemlerinin yetenek kestirimlerine etkisi ile (Wang ve Chang, 2011; Yao, 2012, 2014), ÇBBBT uygulamalarında farklı durdurma kurallarının etkisi üzerinde yoğunlaşmaktadır (Wang, Chang ve Boughton, 2011; Yao, Pommerich, Segall, 2014).

Hem TBBBT hem de ÇBBBT uygulamalarında alt boyut ve toplam puanlarını kestirmede yetenek kestirim yöntemleri doğrudan sonuçları etkilese de, uygun maddelerin seçilmemesi durumunda hiçbir yetenek kestirim yöntemi fonksiyonel olmayacaktır (Reckase, 2009). Maddelerin çok zor, çok kolay ve düşük bilgi verici maddeler arasından seçilmesi yetenek kestirimlerini etkilemektedir. Maddelerin havuzdan seçme işlemindeki kurallar, ÇBBBT uygulamalarında önem kazanmaktadır. Literatürdeki madde seçme yöntemleri θ kestirimde kullanılacak bazı kritik değerleri maksimize ya da minimize

etme ilkesine dayanmaktadır. Madde seçme yöntemleri de kritik değerleri tanımlama konusunda birbirinden farklılaşmaktadır.

ÇBBBT Madde Seçme Yöntemleri

Fisher Bilgi Matrisinin Determinantının Artırtılması-Hacim Yöntemi (Vol)

Segall (1996) önceki araştırmalarda (Bloxom & Vale, 1987; Tam, 1992) çok boyutlu madde tepki kuramı çerçevesinde yeteneğin ortak dağılımına ait önsel bilgilerin kullanılmadığı için sonuçların geçerli sayılamayacağını belirtmiştir. Segall'e göre ÇBBBT her bir alt boyuttan belirli sayıda madde seçme yerine her bir alt boyutun özelliğini etkili şekilde ortaya çıkaracak madde seçme prosedürleri sağlayabilmektedir. ÇBBBT uygulamalarında ayrıca boyutlar arasındaki ilişkiler dikkate alındığından madde seçme prosedürlerinin etkililiği daha da artırılabilmektedir. Segall (1996) Bayes modellemesine dayalı, yeteneğin ortak dağılımına ait önsel bilgileri de dikkate alan bir madde seçme yöntemi önermiştir. Yöntem, bireylerin alt boyut yeteneklerini ÇBMTK modelleri yardımıyla seçilen maddelerden $(k-1)$ kestirmektedir. Bu maddelerden elde edilen bilgi $(I_{k-1}(\theta^{k-1})$ önsel dağılım olarak kullanılmakta ve bir sonraki maddenin (k) seçiminde kullanılmaktadır. Bu sayede yetenek kestirimlerinin $\left(\theta^{k-1}\right)$ doğruluğunun artırıldığı belirtilmektedir. Sonsal bilgi dağılımının determinantını maksimize eden eşitlik denklem 1'de gösterilmiştir.

$$W = \left|I_{k-1}\left(\theta^{k-1}\right) + I_k\left(\theta^{k-1}\right) + \sum^{-1}\right| \quad (1)$$

Madde havuzundaki her i maddesi için, hacim ya da bilgi fonksiyonunun determinantı denklem 2'deki eşitlik ile hesaplanabilmektedir (Yao, 2012).

$$W_m = \left|I_{k-1}\left(\theta^{k-1}\right) + \frac{(P_{i1}-\beta_{3i})^2(1-P_{i1})}{P_{i1}(1-\beta_{ik})^2}\beta_{2i}x\hat{\beta}_{2i} + \sum^{-1}\right| \quad (2)$$

Kullback-Leibler

Kullback-Leibler (KL) bilgisinin TBBBT uygulamalarında ilk olarak Chang ve Ying (1996) tarafından kullanılmıştır. Veldkamp ve van der Linden (2002) KL bilgisini gölge test yöntemi (shadow test method) kullanarak çok boyutlu yapıya uyarlamışlardır. KL madde seçme yöntemi gerçek yetenek (θ_0) ile kestirilen yetenek (θ) arasındaki iki olabilirlik arasındaki uzaklığı ölçmektedir. M3PL model için KL bilgisi denklem 3 ile gösterilmiştir.

$$K_i(\theta, \theta_0) = P_i(\theta_0)Ln\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + [1 - P_i(\theta_0)]Ln\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right] \quad (3)$$

Denklem 3'te $i = (1,2,\ldots,N)$ madde havuzundaki N sayıdaki maddeyi belirtmektedir. θ_o değeri, θ_o bilinmediğinde ve θ tanımlanmadığı durumda, sonsal beklenen KL bilgisine göre kestirilmektedir. θ sonsal dağılımının yoğunluğu $f(\theta\,|u_{i_1},\ldots,u_{i_{k-1}})$ ile tanımlanmaktadır ve uygulanan $k-1$ sayıdaki maddenin bir fonksiyonudur. Sonsal beklenen KL bilgisi kullanılarak $\hat{\theta}^{k-1}$ kestirimi denklem 4 kullanılarak hesaplanmaktadır.

$$K_i^B(\hat{\theta}^{k-1}) \equiv \int_\theta K_i\left(\theta, \hat{\theta}^{k-1}\right) f(\theta\,|u_{i_1},\ldots,u_{i_{k-1}})\partial\theta \quad (4)$$

Çok boyutluluk açısından bakıldığında KL denklemindeki θ ve θ_0 değerleri ÇBMTK modelinde skaler yerine vektörel olarak ifade edilmektedir.

ÇBBBT uygulamalarında KL madde seçme yönteminin kullanılmasının iki temel nedeni bulunmaktadır. İlk olarak, tek boyutlu KL madde seçme yönteminde gerçek θ değerlerinin kestiriminde Fisher bilgisinden daha başarılıdır. Ayrıca KL bilgisi önsel dağılımları kullandığından gerçek ve kestirilen yetenekleri geçerli bir şekilde ayırt etmektedir. KL yöntemi, Fisher yönteminin aksine θ ve θ_0 değerlerinin birer fonksiyonu olarak ifade edilebilir ve yetenek seviyelerinin birbirine yakın olmasını gerektirmez (Chang ve Ying, 1996; Veldkamp ve van der Linden, 2002).

Minimum Hata Varyansı Kriteri

Hata varyanslarının doğrusal birleşimlerinin minimize eden (V1) madde seçme yöntemi, eşit ağırlıklandırılmış boyutlardan elde edilen toplam puanlara en düşük hata varyansı veren maddeyi seçmektedir. Bu yöntem toplam puanların doğruluğunu artırmak için van der Linden (1999) tarafından ortaya atılmıştır.

van der Linden (1999) çok değişkenli bilgi matrisinin yerine asimptotik varyans-kovaryans matrisinin kullanılmasındaki amacın madde seçme yöntemini çok boyutlu MTK'ya göre uyarlamak olduğunu belirtmiştir. Bireyselleştirilmiş test uygulamalarında kestirilen yetenek, her bir madde seçiminden sonra elde edilen yeteneklerin $(\lambda(\theta_1, \dots, \theta_m);\ \lambda = \lambda_j = (\lambda_1, \lambda_2, \dots, \lambda_m);\ \lambda_j \geq 0)$ doğrusal kombinasyonlarına eşittir. Ağırlıklandırmanın (λ) değeri testin amacına göre değişmekte ve bu değer, BBT prosedürlerini ve yetenek kestirimlerini değiştirebilmektedir.

İlk olarak işlem MAP yöntemi kullanarak yetenek kestirimi ile başlamaktadır. Yerel bağımsızlık varsayımından dolayı $\widehat{\lambda'\theta} = \lambda'\theta$ olarak ifade edilebilir ve θ olabilirlik fonksiyonu denklem 5 yardımı ile hesaplanabilmektedir.

$$g\left(\theta \middle| u_{i_1}, u_{i_2}, \dots, u_{i_{k-1}}\right) = \frac{L\left(\theta \middle| u_{i_1}, u_{i_2}, \dots, u_{i_{k-1}}\right) g(\theta)}{\int L\left(\theta \middle| u_{i_1}, u_{i_2}, \dots, u_{i_{k-1}}\right) g(\theta) d\theta} \qquad (5)$$

Bireyselleştirilmiş test algoritmasında herhangi iki yetenek değişkenine ait madde seçme prosedürü için senaryo belirtilen şekildedir: $k-1$ madde seçilmiş olsun. $S_k = (i_1, i_2, \dots, i_{k-1})$ seçilen maddeleri; $R_k = (1,2, \dots, i)/S_k$ ise reddedilen maddeleri göstermektedir. Bireyselleştirilmiş test algoritması kullanılarak $k-1$ sayıda madde uygulandıktan sonra k maddesi denklem 6'daki kritere göre seçilmektedir.

$$min_{R_k}\left[Var(\lambda\hat{\theta}_1^k + (1-\lambda)\hat{\theta}_2^k | \hat{\theta}_1^{k-1}, \hat{\theta}_2^{k-1}\right] \qquad (6)$$

Denklem 6'da $\hat{\theta}_1^k$ ve $\hat{\theta}_2^k$, θ_1 ve θ_2 değerlerinin kestirimlerini göstermektedir. Denklem incelendiğinde madde $\lambda\hat{\theta}_1^k + (1-\lambda)\hat{\theta}_2^k$ varyansını minimize edecek şekilde seçilmektedir. R_k kümesindeki maddeleri seçmek için, final denklemi madde parametrelerini $(a_{i1}, a_{i2}\ ve\ d_i)$ ve olasılık $[P_i(\theta_1, \theta_2)]$ değerlerini içermektedir. Özetlenecek olursa, algoritma ilk olarak testin amacını yansıttığı düşünülen ağırlıklandırılmış değerlerini (λ) seçmektedir. Ağırlıklandırılmış deneysel veya seçilen problemden elde edilmiş olabilmektedir. En son adımda ise yetenek parametreleri MAP eşitlikleri kullanılarak kestirilmektedir.

V1 yöntemini kullanmanın bir takım avantajları vardır (van der Linden, 1999; Yao, 2012). Bazı durumlarda madde havuzu birden çok yeteneği ölçecek şekilde desenlenmiştir. Bu yetenekler arasında ilişkiler olabileceği gibi herhangi bir ilişkiye rastlanmıyor olabilir. Bu durumda bireysel yetenekler ağırlıklandırma seçimini etkileyebilmektedir. Bu bakımdan V1 yöntemi alt boyut puanlarının doğrusal kombinasyonlarını düzenleyerek toplam puanların daha doğru kestirilmesini sağlamaktadır

Literatürde toplam puanları rapor etmede çok boyutlu madde seçme yöntemlerinin karşılaştırılmasına ilişkin çalışmalar olsa da (Wang ve Chang, 2011; Yao, 2012, 2013), çok boyutlu test deseninin yapısına göre nasıl performans gösterdiğine ilişkin bir çalışmaya rastlanmamıştır.

Araştırmanın Amacı

Çok boyutlu bireyselleştirilmiş bilgisayarlı test (ÇBBBT) uygulamalarının tek boyutlu bireyselleştirilmiş bilgisayarlı test (TBBBT) uygulamalarına göre birtakım üstünlükleri bulunmaktadır. ÇBMTK test yapılarında boyutlar arasında ilişkiler mevcuttur ve bir madde birden çok boyuta yük verebilir. Bu sayede her bir maddenin bilgi vericiliği ÇBBBT uygulamalarında daha dikkatle ele alınır ve TBBBT'ye göre daha kararlı sonuçlar elde edilir. Maddelerden elde edilen bilgiler arttığından test uzunluğu da düşmektedir (Segall, 1996). TBBBT uygulamalarında kullanılan kapsam dengeleme kısıtlamaları sonucu bazı alt boyutlar bireyin genel yeteneğine daha az katkı sağlamaktadır. ÇBBBT uygulamaları, kapsam alanlarını korelasyon değerlerini de göz önüne alarak ayrı ayrı ele alır

ve farklı kapsamlardan elde edilen bilgiyi bütün boyutlarla beraber işleme koyar (Segall, 1996; Wang ve Chang, 2011).

İncelenen araştırmalar madde sayısının ve maddeler arası korelasyon değerlerinin alt boyut ve toplam puanları hesaplamada değişebildiğini göstermektedir (Segall, 1996; Wang ve Chang, 2011; Wang, Chang ve Boughton, 2013; Yao, 2012) . Bu sebeple boyutlar arası korelasyon ve boyut başına düşen madde sayısının madde seçme yöntemleri üzerindeki etkisinin nasıl değişeceğinin belirlenmesinin uygulayıcılara önemli bilgiler sağlayacağı düşünülmektedir.

ÇBBBT uygulamalarında toplam puanları hesaplamada madde seçme yöntemlerinin farklı kestirim değerleri sunduğu belirlenmiştir (Yao, 2012, 2013). Bu sebepten madde seçme yöntemlerinin karşılaştırılması ve toplam puanları kestirmede farklı koşullar için en az hata veren yöntemlerin belirlenmesi gerekmektedir. Geniş ölçekli sınavlarda puanların rapor edilmesinin her geçen gün arttığı, sınavı alan bireylere bu puanlar doğrultusunda geri bildirimler verildiği dikkate alındığında ÇBBBT uygulamalarında toplam puanları rapor etmede en az hata içeren koşulların belirlenmesinin önemli olduğu düşünülmektedir. Bu araştırmada, PISA 2012 Türkiye örnekleminden elde edilen veriler kullanılarak çok boyutlu test yapılarında toplam puanları belirlemede madde seçme yöntemlerinin farklı koşullar altındaki performanslarını ortaya çıkarmak amaçlanmıştır.

Problem Cümlesi

Basit, düşük ve yüksek karmaşık yapıdaki testlerde, boyutlar arası korelasyonun ve boyut başına düşen madde sayısının madde seçme yöntemlerinin hata, mutlak yanlılık ve korelasyon değerlerine etkisi nasıldır?

Alt Problemler

1. Basit Yapılı (BY) test deseninde çok boyutlu madde seçme yöntemlerinin hatası, mutlak yanlılığı ve korelasyon değerleri boyutlar arası korelasyona ve boyut başına düşen madde sayısına göre toplam yetenek puanları için nasıl değişmektedir?
2. Düşük Karmaşık Yapılı (DKY) test deseninde çok boyutlu madde seçme yöntemlerinin hatası, mutlak yanlılığı ve korelasyon değerleri boyutlar arası korelasyona ve boyut başına düşen madde sayısına göre toplam yetenek puanları için nasıl değişmektedir?
3. Yüksek Karmaşık Yapılı (YKY) test deseninde çok boyutlu madde seçme yöntemlerinin hatası, mutlak yanlılığı ve korelasyon değerleri boyutlar arası korelasyona ve boyut başına düşen madde sayısına göre toplam yetenek puanları için nasıl değişmektedir?

YÖNTEM

Araştırmada var olan yöntem ve tekniklerin gerçek veri üzerinden performanslarının karşılaştırılması amaçlandığından araştırma nicel karşılaştırma araştırmasıdır.

Araştırmanın Deseni

Araştırmada test deseni, boyutlar arası korelasyon değeri, boyut başına düşen madde sayısı ve madde seçme yöntemleri olmak üzere dört farklı durum manipüle edilmiştir. Manipülasyonların sonucunda 3x3x2x3 olmak üzere toplam 54 deneysel koşul çapraz olarak test edilmiştir.

Tablo 1. Araştırma Deseni

Test Deseni	Boyutlar arası korelasyon	Boyut Başına Düşen Madde Sayısı	Madde Seçme Yöntemi
Basit Yapı (BY) Düşük Karmaşık Yapı (DKY) Yüksek Karmaşık Yapı (YKY	ρ=0.2 ρ=0.5 ρ=0.8	Kısa Test (n=10) Uzun Test (n=15)	Kullback-Leibler (KL) Hacim (Vol) Minimum Hata Varyansı (V1)

Verilerin Üretilmesi

Araştırmada kullanılan madde parametrelerinin üretilmesinde PISA 2012 Türkiye verisinden yararlanılmıştır. Eğim, güçlük ve düşük asimptot parametreleri a, b ve c, 3 parametreli lojistik model kullanılarak a ~ LN{0, 0.2), b ~ N(0, 1), ve c ~ Beta{6,16) koşullarını sağlayacak şekilde üretilmiştir. Ampirik olarak elde edilen madde parametresi değerlerine bağlı kalmak koşuluyla a ve b parametreleri [0.5, 1.5] ve [-2, 2] arasında değerler alıncaya kadar yeniden üretilmiştir. Düşük asimptot değeri olan c-parametresi 0.15 değerine sabitlenmiştir.

Test Desenlerinin Oluşturulması

Araştırmada çok boyutlu test desenleri, tek boyutlu olarak üretilen *a*-parametrelerinden yararlanılarak ve

$$a_j = MDISC = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

maddenin çok boyutlu ayırt ediciliği (ÇBAE-MDISC) formülü kullanılarak belirlenmiştir. MDISC değeri çok boyutlu ayırt ediciliğin tek boyutlu halidir (Ackerman, Gierl ve Walker, 2003).

Araştırmada çok boyutlu test yapılarından basit yapı, düşük karmaşık yapı ve yüksek karmaşık yapı olmak üzere üç farklı test deseni çok boyutlu ayırt edicilik değerlerinin sabit tutulması koşuluyla *a*-parametresi değerlerinin boyutlara dağıtılmasıyla belirlenmiştir.

Yetenek Parametrelerinin Üretilmesi

BY, DKY ve YKY test desenleri için 1000 birey ve 3 alt boyuttan oluşan 1000x3 gerçek yetenek parametreleri matrisi, çok değişkenli normal dağılıma göre $\theta_i = MVN(0, \sum)$) varyans-kovariyans matrisleri kullanılarak üretilmiştir. Simülasyon sonucu elde edilen madde ve yetenek parametreleri kullanılarak, boyutlar arası korelasyon değerleri ile birlikte, telafi edici çok boyutlu MTK modeline göre cevap matrisleri MIRTGEN 3.0 (Luecht, 2004) programı kullanılarak üretilmiştir. Cevap matrisleri üretilirken çok boyutlu 3 parametreli lojistik model kullanılmıştır (M3PL). Maddelerin kalibrasyonunda MML yöntemine dayanan Bock ve Aitkin Expectation-Maximization (BAEM) algoritması (Bock ve Aitkin, 1981) kullanılmıştır. BAEM algoritması EM algoritmasından farklı olarak her bir madde parametresinin log-olabilirlik değerlerinin türevlerini sadece o madde parametrelerine bağlı olarak hesaplanmaktadır.

Verilerin Analizi

ÇBBBT analizlerinde yetenek kestirimleri ve madde seçme yöntemleri her boyuta aynı anda uygulanmıştır. Başlangıç maddesi $\theta_{başlangıç} = \{\theta_1, \theta_2, \theta_3\} = \{0,0,0\}$ koşulunu sağlayacak şekilde seçilmiştir. Test uzunlukları madde havuzlarının uzunluklarına eşit olacak şekilde her bir boyut için toplamda 30 ve 45 olarak belirlenmiştir ve sabit uzunluklu sonlandırma kuralı uygulanmıştır.

Araştırmada ÇBBBT alt boyut ve toplam yetenek puanları MAP yöntemi kullanılarak 100 iterasyonun ortalaması hesaplanarak kestirilmiştir. ÇBBBT toplam yetenek puanlarının kestirimi SimuMCAT (Yao, 2011) programı kullanılarak hesaplanmıştır.

Değerlendirme Kriteri

Her bir ÇBBBT koşulu için, toplam yetenek parametrelerinin kesinliğini belirlemede gerçek ve kestirilen yetenek puanları arasındaki korelasyon katsayısı ($r = \frac{\sigma_{\hat{\theta}\theta}}{\sigma_{\hat{\theta}}\sigma_{\theta}}$), hata kareleri ortalamasının karekökü ($RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}$) ve mutlak yanlılık ($ABSBIAS = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta}_i - \theta_i\right|$) kullanılmıştır.

BULGULAR

Gerçek yetenek parametreleri basit, düşük karmaşık ve yüksek karmaşık test desenlerine, boyutlar arasındaki korelasyona ve her bir boyuta düşen soru sayısına göre çok boyutlu normal dağılım kullanılarak üretilmiştir. Üretilen parametrelerin kesinliğini kontrol etmek amacıyla boyutlar arası korelasyon değerleri, ortalama ve standart sapma değerleri her bir koşul için hesaplanmıştır. Tablo 2 MIRTGEN (Luecht, 2004) programı kullanılarak üretilen gerçek yetenek parametrelerinin sonuçlarını özetlemektedir.

Tablo 2. Gerçek Yetenek Puanlarına Ait Ortalama, Standart Sapma ve Boyutlar arası Korelasyon Değerleri

			Altboyut 1		Altboyut 2		Altboyut 3				
Madde Sayısı*	Test Deseni	Korelasyon**	Ort	Ss	Ort	Ss	Ort	Ss	ρ'_{12}	ρ'_{13}	ρ'_{23}
10 Madde	BY	$\rho = 0.2$	.006	.99	-.067	.99	-.059	.99	0.191	0.229	0.247
		$\rho = 0.5$	.032	.97	.047	.96	-.012	.98	0.462	0.473	0.443
		$\rho = 0.8$	-.048	1.02	-.058	1.01	-.058	1.02	0.814	0.827	0.806
	DKY	$\rho = 0.2$	.013	.97	-.034	1.01	-.057	.99	0.268	0.192	0.276
		$\rho = 0.5$	.000	.96	.028	1.02	-.014	.98	0.508	0.548	0.528
		$\rho = 0.8$	-.005	.98	-.018	.98	.005	.98	0.809	0.803	0.792
	YKY	$\rho = 0.2$	.084	1.01	.032	1.01	.051	.98	0.148	0.221	0.227
		$\rho = 0.5$	.035	1.06	.034	1.01	-.001	.99	0.543	0.539	0.501
		$\rho = 0.8$	.034	1.00	.044	1.01	.005	1.01	0.803	0.802	0.781
15 Madde	BY	$\rho = 0.2$	.013	.99	-.009	1.01	.009	.99	0.185	0.185	0.231
		$\rho = 0.5$	.006	1.01	-.026	1.04	.012	1.02	0.526	0.490	0.495
		$\rho = 0.8$	-.004	.99	-.021	1.03	.002	1.01	0.798	0.789	0.814
	DKY	$\rho = 0.2$	-.008	.98	.000	1.01	-.014	.94	0.209	0.173	0.194
		$\rho = 0.5$	-.035	.96	-.012	1.01	-.048	1.04	0.505	0.517	0.502
		$\rho = 0.8$	-.020	1.03	-.013	1.02	.009	1.03	0.798	0.813	0.816
	YKY	$\rho = 0.2$	-.024	.99	-.034	1.01	-.052	.99	0.170	0.228	0.196
		$\rho = 0.5$	-.061	.98	-.027	.99	-.001	.96	0.492	0.516	0.498
		$\rho = 0.8$	-.015	1.01	-.021	.97	-.018	1.01	0.793	0.794	0.802

**Boyut başına düşen madde sayısı, **Boyutlar arası korelasyon*

Tablo 2 incelendiğinde üretilen yetenek parametrelerine ait korelasyon değerleri varsayılan (hipotetik) korelasyon değerlerine benzer olarak elde edilmiştir. Her bir alt boyuta ait üretilen gerçek yetenek puanlarına ait ortalamalar -0.067 ile 0.084 arasında, standart sapma değerleri ise 0.94 ile 1.06 arasında

değişmektedir. Alt boyutlara ait gerçek puanlar, ortalaması 1, standart sapması 0 olan çok boyutlu normal dağılıma yakın değerlerde elde edilmiştir.

Tablo 3. Test Desenlerine Ait Mutlak Yanlılık, Hata ve Korelasyon Değerleri

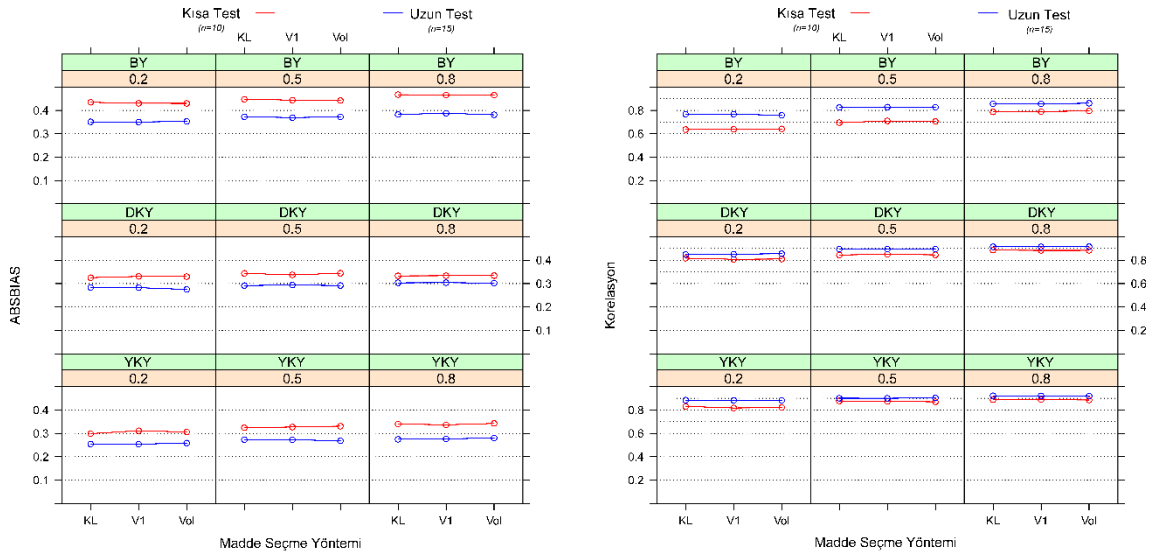
	BY			DKY			YKY		
	KL	V1	Vol	KL	V1	Vol	KL	V1	Vol
ABSBIAS									
$\rho = 0.2$, N=10	.328	.323	.327	.244	.251	.246	.232	.248	.240
$\rho = 0.5$, N=10	.358	.349	.355	.271	.268	.274	.250	.251	.259
$\rho = 0.8$, N=10	.359	.345	.354	.269	.258	.263	.263	.261	.270
$\rho = 0.2$, N=15	.307	.302	.313	.191	.190	.185	.185	.188	.189
$\rho = 0.5$, N=15	.290	.286	.289	.206	.202	.209	.204	.204	.200
$\rho = 0.8$, N=15	.302	.304	.306	.213	.213	.209	.207	.206	.210
RMSE									
$\rho = 0.2$, N=10	.014	.014	.014	.010	.010	.010	.009	.010	.010
$\rho = 0.5$, N=10	.014	.014	.014	.011	.011	.011	.010	.010	.010
$\rho = 0.8$, N=10	.015	.015	.015	.011	.011	.011	.011	.011	.011
$\rho = 0.2$, N=15	.011	.011	.011	.009	.009	.009	.008	.008	.008
$\rho = 0.5$, N=15	.012	.012	.012	.009	.009	.009	.009	.009	.008
$\rho = 0.8$, N=15	.012	.012	.012	.010	.010	.010	.009	.009	.009
Korelasyon									
$\rho = 0.2$, N=10	.64	.64	.64	.82	.81	.81	.83	.83	.82
$\rho = 0.5$, N=10	.70	.71	.70	.84	.85	.85	.88	.87	.87
$\rho = 0.8$, N=10	.79	.79	.79	.89	.88	.88	.89	.89	.88
$\rho = 0.2$, N=15	.77	.77	.76	.85	.85	.86	.88	.88	.88
$\rho = 0.5$, N=15	.82	.83	.83	.90	.90	.90	.90	.90	.90
$\rho = 0.8$, N=15	.86	.86	.86	.91	.91	.92	.92	.92	.92

Tablo 3 incelendiğinde basit yapılı bir testte testin katı çok boyutluluk (ρ=0.2) özeliği gösterdiği ve boyut başına düşen madde sayısının n=10 olduğu durumda toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.328, 0.323 ve 0.327 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri de 0.014; korelasyon değerleri ise 0.64 olarak hesaplanmıştır. Boyut başına düşen madde sayısının 10 olduğu kısa testlerde, toplam yetenek puanlarında çok boyutluluk değerlerine göre ABSBIAS ve RMSE değerlerinde kısmi artışlar gözlenmiştir. Boyut başına düşen madde sayısının 10 olduğu ve çok boyutluluğun etkisinin azaldığı durumda, yani boyutlar arası korelasyonun 0.2'den 0.8'e çıktığı durumda, korelasyon ile birlikte RMSE ve ABSBIAS değerleri de artmıştır. Testin çok boyutluluğunun azaldığı (ρ=0.8) ve boyut başına düşen madde sayısının n=10 olduğu durumda toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.359, 0.345 ve 0.354 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri 0.015; korelasyon değerleri ise 0.79 olarak hesaplanmıştır.

Tablo 3 incelendiğinde test yapısı karmaşıklaşmaya başladıkça mutlak yanlılık ve hata değerlerinde azalma olduğu görülmüştür. Test yüksek karmaşık yapıda olduğu durumda en düşük mutlak yanılık değerleri Tablo 3'te belirtilmiştir. Boyut başına düşen madde sayısının 10 olduğu durumda, düşük yapılı testteki mutlak yanlılık ve hata değerleri basit yapıdakiler ile benzerlik göstermiş, boyutlar arası korelasyon değerinin 0,2'den 0,5'e yükselmesi ile artmış; 0,5'ten 0,8'e çıkmasıyla azalmıştır. Yüksek karmaşık yapılı testlerde ise boyutlar arası korelasyon değeri arttıkça mutlak yanlılık ve hata değerleri sürekli olarak artmıştır. Boyut başına düşen madde sayısının 10 olduğu durumda en düşük mutlak yanlılık ve hata değerleri yüksek karmaşık yapıdaki test deseninde görülmüştür. Ayrıca gerçek puanlar ile kestirilmiş puanlar arasındaki korelasyon değerlerinin en yüksek değer aldığı test deseni de yüksek karmaşık yapıdaki test desenidir.

Testin uzunluğu artırıldığında (n=15) beklenildiği gibi yetenek puanları daha az hata ve mutlak yanlılık ile kestirilmiştir. Testin katı çok boyutluluk (ρ=0.2) özeliği gösterdiği ve boyut başına düşen madde sayısının n=15 olduğu durumda toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.307, 0.302 ve 0.313 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri 0.011; korelasyon değerleri ise sırasıyla 0.77, 0.77 ve 0.76 olarak hesaplanmıştır. Bu bulguların sonucuna bakıldığında tüm test desenleri için testin katı çok boyutluluk (ρ=0.2) özelliği gösterdiği durumda, boyut başına düşen madde sayısının artırılması, toplam yetenek puanlarının kestirimini olumlu olarak etkileyecektir denilebilir.

Basit yapıdaki bir testte kısa testlerin aksine uzun testlerde boyutlar arası korelasyon değeri arttığında RMSE ve ABSBIAS değerlerinde bir azalma görülmüştür. Testin çok boyutluluğunun azaldığı (ρ=0.8) ve boyut başına düşen madde sayısının n=15 olduğu durumda ise toplam yetenek puanlarının KL, V1 ve Vol değerleri için ortalama mutlak yanlılık (ABSBIAS) değerleri sırasıyla 0.302, 0.304 ve 0.306 olarak hesaplanmıştır. Her üç madde seçme yöntemine ait RMSE değerleri de 0.012; korelasyon değerleri ise 0.86 olarak hesaplanmıştır. Test yapısı karmaşıklaşmaya başladıkça mutlak yanlılık ve hata değerleri boyutlar arası korelasyonun artışı ile doğru orantılı olarak artmıştır. Toplam test puanlarına ait en düşük mutlak yanlılık ve hata değerleri testin yüksek karmaşık yapıda olduğu durumda görülmüştür. Ayrıca gerçek puanlar ile kestirilmiş puanlar arasındaki korelasyon değerlerinin en yüksek değer aldığı test deseni de yüksek karmaşık yapıdaki test desenidir (r=0.92).



Şekil 1. Madde Seçme Yöntemlerinin Mutlak Yanlılık ve Korelasyon Değerleri

Şekil 1'de madde seçme yöntemlerinin test desenine, boyutlararası korelasyon değerine ve boyut başına düşen madde sayısına göre mutlak yanlılık ve korelasyon değerleri gösterilmiştir. Şekil 1 incelendiğinde, boyular arası korelasyonun düşük (ρ=0.2) ve boyut başına düşen madde sayısının n=10 olduğu durumda basit yapıdaki testlerde V1; test karmaşıklaştığı durumlarda ise KL madde seçme yöntemi en iyi performansı göstermiştir. Boyular arası korelasyonun yüksek (ρ=0.8) ve boyut başına düşen madde sayısının n=10 olduğu durumda tüm test desenlerinde V1 madde seçme yöntemi en iyi performansı göstermiştir.

Boyutlar arası korelasyonun düşük (ρ=0.2) ve boyut başına düşen madde sayısının n=15 olduğu durumda basit yapıda V1, düşük karmaşık yapıda Vol ve yüksek karmaşık yapıda ise KL madde seçme yöntemi en iyi performansı göstermiştir. Boyutlar arası korelasyonun yüksek (ρ=0.8) ve boyut başına düşen madde sayısının n=15 olduğu durumda basit yapıda KL, düşük karmaşık yapıda Vol ve yüksek karmaşık yapıda ise V1 madde seçme yöntemi en iyi performansı göstermiştir.

SONUÇLAR ve TARTIŞMA

Araştırmada üç farklı madde seçme yönteminin performansı karşılaştırılmıştır. Sonuçlar incelendiğinde önce V1 daha sonra da KL yönteminin daha iyi performans gösterdiği belirlenmiştir. Yao (2012) her bir boyuttan belirli sayıda madde seçilmesi durumunda KL ve V1 madde seçme yöntemlerinin alt boyut ve toplam puanları kestirmede daha iyi sonuçlar vereceğini belirtmiştir. Bu bakımdan sonuçlar literatürdeki araştırmalarla örtüşmektedir. Ayrıca V1 yönteminin daha iyi performans göstermesindeki en büyük neden ÇBBBT işlemlerinde ÇBMTK uygulamaları sonucu elde edilen boyutlara arası teorik ağırlıklandırmalar yerine Tablo 1'de belirtilen kestirilen ağırlıkların kullanılmasıdır.

Toplam puanların kestirilmesinde BY test deseninde V1 madde seçme yöntemi daha ağırlıklı olarak seçilmekte iken, test yapısının karmaşıklaştığı durumda KL ve Vol yöntemleri de iyi performans göstermiştir. Özellikle Vol yöntemi her bir boyuttan elde edilen bilgiyi eşit dağıtmaya çalışmaktadır (Yao, 2012). Test karmaşık yapıya doğru gittikçe bilginin boyutlara belirli oranlarda dağıldığı bilinmektedir. Bu bakımdan test karmaşıklaştıkça Vol madde seçme yönteminin belirli koşullarda daha iyi performans vermesi beklenen bir durumdur. KL diğer iki yöntemden farklıdır ve MDISC değerine göre madde seçmektedir. KL olabilirlik fonksiyonuna göre maddeleri seçtiğinden dolayı, olabilirlik fonksiyonları birbirinden uzak olan yetenek dağılımlarında çok az sayıda iyi maddeleri seçme eğilimindedir. Araştırmada sabit sayıda soru sorulduğundan ve boyutlardaki sorular önceden belirlendiğinden dolayı KL madde seçme yönteminin avantajlarının tam olarak yansıtılmadığı düşünülmektedir.

Genel çerçevede bakıldığında madde sayısı arttığında korelasyonların arttığı ve mutlak yanlılık ve hata değerlerinin ise azaldığı belirlenmiştir. Toplam yetenek puanlarında boyut başına düşen madde sayısının 15 olduğu bir teste ait mutlak yanlılık değeri ise tüm koşullarda boyut başına düşen madde sayısının 10 olduğu bir testin mutlak yanlılık değerlerinden düşük olarak kestirilmiştir. Elde edilen bu bulgular literatürdeki çalışmaları destekler niteliktedir (Lee, 2014; Su, 2016; Yao, 2014; Yao, Pommerich ve Segall, 2014).

Tüm test desenlerinde, madde seçme yöntemi fark etmeden, boyutlar arası korelasyon değeri arttığında boyut başına düşen madde sayısına göre toplam yetenek puanlarına ait mutlak yanlılık değerlerinde farklılaşmalar görülmüştür. Basit ve düşük karmaşık yapı test deseninde test katı çok boyutluluk özelliğinden tek boyutluluğa yaklaştığı durumlarda kısa test için mutlak yanlılık değerleri önce artmış daha sonra V1 ve Vol kestirimleri için azalmıştır. Uzun testlerde ise her üç madde seçme yönteminde önce bir azalış daha sonra da bir artış görülmüştür. Ancak uzun testlerde yapı karmaşıklaşmaya başlayınca boyutlar arası korelasyon arttığında her üç madde seçme yöntemine göre mutlak yanlılık değerleri düzgün şekilde artmıştır. Yüksek karmaşık yapılarda ise hem kısa hem uzun testler madde seçme yöntemleri farketmeksizin, boyutlar arası korelasyon değerleri arttıkça daha fazla mutlak yanlılık değeri üretmişlerdir.

Araştırmada ayrıca tüm madde seçme yöntemlerinde testin karmaşıklığı arttığında yanlılık ve hata değerlerinin azaldığı görülmüştür. Testin karmaşıklığı arttıkça, boyutlar arası korelasyon değerinin 0.2 olduğu durumda toplam puanların raporlanmasındaki korelasyon değerleri arasındaki fark fazla iken, boyutlar arasındaki korelasyonun 0.8 olması durumunda ise birbirine çok yakın değerler elde etmişlerdir. Bu bakımdan testin hem karmaşıklığının artırılması hem de boyutlar arası korelasyonun artırılması toplam puanları rapor etmede mutlak yanlılıklarda önemli bir durumdur. Bu durumun telafi edici modellerin etkisinden olduğu düşünülmektedir. Elde edilen bulgular literatürdeki çalışmaları destekler niteliktedir (Su, 2016).

Araştırma PISA 2012 Türkiye verilerine dayalı bir simülasyon çalışması olarak ele alınmış ve gerçek test koşullarına en yakın durumlar manipüle edilmiştir. Araştırma gerçek ÇBBBT uygulamaları üzerinden yapılarak madde ve yetenek dağılımına ilişkin sonuçlar karşılaştırılabilir. Ayrıca, ÇBBBT yöntemlerinin etkililiklerin ortaya konmasında madde havuzlarının önemi büyüktür. Bu bakımdan benzer koşullar daha büyük çok boyutlu madde havuzlarında test edilebilir ve madde seçme yöntemlerinin performansı karşılaştırılabilir.

KAYNAKÇA

Ackerman, T. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255-278. doi: 10.1207/s15324818ame0704_1

Ackerman, T. A., Gierl, M. J., & Walker, C. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37–53. doi: 10.1111/j.1745-3992.2003.tb00136.x

Adams R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. doi: 10.1177/0146621697211001

Bloxom, B., & Vale, C. D. (1987). *Multidimensional adaptive testing: An approximate procedure for updating*. Paper presented at the annual meeting of the psychometric society. Montreal, Canada.

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

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation). University of Minnesota, USA.

Fan, M., & Hsu, Y. (1996). *Multidimensional computer adaptive testing*. Paper presented at the annual meeting of the American Educational Testing Association, New York City, NY.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement, 34*(1), 10–26. doi: 10.1177/0146621609336112

Lee, M. (2014). *Application of higher-order IRT models and hierarchical IRT models to computerized adaptive testing* (Unpublished doctoral dissertation). University of California, Los Angeles.

Liu, F. (2015). *Comparisons of subscoring methods in computerized adaptive testing: A simulation study* (Unpublished doctoral dissertation). University of North Carolina Greensboro, North Carolina, USA.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404. doi: 10.1177/014662169602000406

Luecht, R. M. (2004). *MIRTGEN 2.0 Manual*. Department of Educational Research Methodology, University of North Carolina at Greensboro, Greensboro, NC.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*(3), 279-293. doi: 10.1177/014662169201600308

Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Luo, X. (2013). *The optimal design of the dual-purpose test* (Unpublished doctoral dissertation). University of North Carolina Greensboro, North Carolina, USA.

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp.77-101). New York: Springer.

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*(2), 273–296. doi: 10.1007/s11336-008-9097-5

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354. doi: 10.1007/BF02294343

Su, Y. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement, 40(*5), 346-360. doi: 10.1177/0146621616639305

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait* (Unpublished doctoral dissertation). Columbia University.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398–412. doi: 10.3102/10769986024004398

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*(4), 575–588. doi: 10.1007/BF02295132

Wang, C., & Chang, HH. (2011). Item selection in multidimensional computerized adaptive tests: Gaining information from different angles. *Psychometrika, 76*(3), 363-384. doi: 10.1007/s11336-011-9215-7

Wang, C., Chang, HH., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika, 76*(1), 13-39. doi: 10.1007/s11336-010-9186-0

Wang, C., Chang, HH., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*, 99-122. doi: 10.1177/0146621612463422

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28,* 295-316. doi: 10.1177/0146621604265938

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9,* 116-136. doi: 10.1037/1082-989X.9.1.116

Yao, L. (2011). *simuMCAT: Simulation of multidimensional computer adaptive testing* [Computer software]. Monterey: Defense Manpower Data Center.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika, 77,* 495-523. doi: 10.1007/s11336-012-9265-5

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37,* 3-23. doi: 10.1177/0146621612455687

Yao, L. (2014). Multidimensional CAT item selection procedures with item exposure control and content constraints. *Journal of Educational Measurement, 51,* 18-38. doi: 10.1111/jedm.12032

Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests containing polytomous items. *Journal of Educational Measurement, 46,* 177–197. doi: 10.1111/j.1745-3984.2009.00076.x

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet precise, screening test. *Applied Psychological Measurement, 38*(8), 614-631. doi: 10.1177/0146621614541514

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36,* 375–398. doi: 10.1177/0146621612445904

EXTENDED ABSTRACT

Introduction

A test can be designed for many purposes, including the ranking of people along a continuum or providing diagnostic value about examinees. However, a very common problem that often arises is the reporting the overall scores when items are designed for multidimensional purposes. Multidimensional computer adaptive testing (MCAT) is capable of measuring multiple dimensions efficiently by using multidimensional IRT (MIRT) applications.

Maximizing the determinant of the fisher information matrix (Vol), Minimizing the error variance of the linear combination (V1) and Kullback-Leibler (KL) item selection methods allow sufficient items from each content by incorporating information from several dimensions simultaneously. Volume item estimation method includes prior knowledge of the joint distribution of ability in a Bayesian framework (Segall, 1996). Volume item selection estimates ability of each domain with using MIRT model by using selected items. This information is used as a prior distribution to select next item which is believed to contribute to the precision of ability estimates. KL information measures the distance between two likelihoods at true ability and current ability and it is concluded that KL information is a better indicator discriminating true and estimated ability based on posterior densities and doesn't require ability levels close to each other (Veldkamp and van der Linden, 2002). Also KL information overcomes the attenuation paradox which helps to estimate correct θ values rather than using Fisher information. Minimizing the error variance of the linear combination (V1, Variance) is more effective when item pool is designed to measure more than one abilities whether these domains are correlated or not. It is stated that V1increasses the precision for overall scores (van der Linden, 1999). Method simply selects the item which contributes minimum error variance for the composite score of equal weighted domains

There have been several research studies about MCAT item selection methods to improve the overall ability score estimations accuracy (Wang and Chang, 2011; Yao, 2012, 2013). According to the literature review it has been found that most studies focused on comparing item selection methods in many conditions except for the structure of test design. In contrast with the previous studies, this study employed various test designs (simple and complex) which allow the evaluation of the overall ability score estimations across multiple real test conditions. The purpose of this study is to compare MCAT item selection methods while estimating the overall ability scores in terms of test design, correlations between dimensions and number of items per dimension in MCAT framework. This study also aims

to find the better item selection procedure which produces higher precisions for the composite score estimations. The comparison is performed by examining the absolute bias (ABSBIAS), root mean square error (RMSE) and correlation between true and estimated ability scores.

Method

In this study, four factors were manipulated, namely the test design, number of items per dimension, correlation between dimensions and item selection methods. For each simple structure (SS), complex low structure (CLS) or complex high structure (CHS) design 1000x3 and 1000x45 matrix of true ability parameters were randomly generated from the multivariate normal distribution. Using the generated item and ability parameters, dichotomous item responses were generated by using M3PL compensatory multidimensional IRT model with specified correlations. A three-dimensional item pool was simulated with simple and complex structures. Dimensions correlated at $\rho = 0.2$, 0.5, and 0.8. For item calibration, multidimensional Bock and Aitkin's EM algorithm (BAEM) calibration method were employed. The multidimensional CAT item selection procedures: minimum angle, minimize the error variance of the composite score with the optimized weight, and Kullback–Leibler (KL) information were also examined. MCAT composite ability score accuracy was evaluated using absolute bias (ABSBIAS), correlation and the root mean square error (RMSE) between true and estimated ability scores.

Results and Discussion

The results suggest that the multidimensional test structure, number of item per dimension and correlation between dimensions had significant effect on item selection methods for the overall score estimations. For SS test design it was found that V1 item selection has the lowest absolute bias estimations for both long and short test while estimating overall scores. For CLS test design it was found that KL item selection has the lowest absolute bias estimations for short test and Vol item selection has the lowest absolute bias estimations for long test while estimating overall scores. For CHS test design it was found that KL item selection has the lowest absolute bias estimations for both long and short test while estimating overall scores.

As the model gets complex absolute biases have decreased significantly for overall scores. Based on the findings V1 item selection had the most accurate estimations for the overall scores. As the number of item increased, correlations tend to increase however, absolute bias and errors decreased. As expected, longer test provided more accurate scores. Correlations increased in overall ability score when the complexity of test increased. Results also suggest that, the overall scores were more sensitive to test complexity (trait contamination), multidimensionality and test length. In all conditions, longer test produced the lowest ABSBIAS and RMSE values and higher correlations.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);47-62

Türkiye'deki Öğretmenlerin Karşılaştıkları Mesleki Sorunların İkili Karşılaştırma Yöntemi İle Ölçeklenmesi*

Scaling Professional Problems of Teachers in Turkey with Paired Comparison Method

Yasemin Duygu ESEN** Filiz TEMEL*** Ergül DEMİR****

Öz

Bu çalışmada öğretmenlerin yaşadıkları mesleki sorunları belirlemek ve bu sorunları temsil eden sorun alanlarının önem düzeyini ikili karşılaştırma yöntemi ile ölçeklemek amaçlanmıştır. Çalışma tarama modelinde yürütülmüştür. Araştırmanın örneklemini Türkiye'de 2015-2016 eğitim ve öğretim yılında Milli Eğitim Bakanlığı'na (M.E.B.) bağlı devlet okullarında görev yapmakta olan 484 öğretmen oluşturmaktadır. Veri toplama aracı olarak araştırmacılar tarafından geliştirilmiş olan "Öğretmenlik Mesleği Mesleki Sorunlar Formu" kullanılmıştır. Verilerin analizinde, Thurstone'nun karşılaştırmalı yargılar kanununun III. Hal denklemi ile ölçekleme yöntemi kullanılmıştır. Araştırma sonucunda elde edilen bulgulara göre öğretmenlerin mesleki sorunları, öğretmen yetiştirme sorunları ve öğretmen niteliği, özlük hakları ve ekonomik sorunlar, mesleki saygının azalması, M.E.B. politikaları ile ilgili sorunlar, sendikal faaliyetlerle ilgili sorunlar, iş yükünün fazlalığı, okuldaki idari sorunlar, fiziki koşullar ve alt yapı yetersizliği, veli ile ilgili yaşanan sorunlar, ve öğrenci ile ilgili sorunlar temaları altında toplanmıştır. Öğretmenlere göre en önemli sorun alanı M.E.B. eğitim politikaları olarak görülmüştür. Bunu sırası ile mesleki saygınlığın azalması, fiziki koşullar ve alt yapı yetersizliği, öğrenci ile yaşanan sorunlar, özlük hakları ve ekonomik sorunlar, okuldaki idari sorunlar, öğretmen yetiştirme ve öğretmen niteliği, veli ile ilgili yaşanan sorunlar, iş yükü fazlalığı, sendikal faaliyetlerle ilgili sorunlar takip etmiştir. Öğretmenlerin mesleki sorun alanları kıdem değişkenine göre incelendiğinde 0-10 yıllık öğretmenler mesleki saygınlığı en önemli sorun alanı olarak görürken, 11-45 yıllık öğretmenler MEB politikalarını en önemli sorun alanı olarak görmüştür.

Anahtar Kelimeler: Öğretmen mesleki sorunları, ikili karşılaştırma yöntemi, ölçekleme

Abstract

In this study, teachers' professional problems was investigated and the significance level of them was measured with the paired comparison method. The study was carried out in survey model. The study group consisted of 484 teachers working in public schools which are accredited by Ministry of National Education (MEB) in Turkey. "The Teacher Professional Problems Survey" developed by the researchers was used as a data collection tool. In data analysis , the scaling method with the third conditional equation of Thurstone's law of comparative judgement was used. According to the results of study, the teachers' professional problems include teacher training and the quality of teacher, employee rights and financial problems, decrease of professional reputation, the problems with MEB policies, the problems with union activities, workload, the problems with administration in school, physical conditions and the lack of infrastructure, the problems with parents, the problems with students. According to teachers, the most significant problem is MEB educational policies. This is followed by decrease of professional reputation, physical conditions and the lack of infrastructure, the problems with students, employee rights and financial problems, the problems with administration in school, teacher training and the quality of teacher, the problems with parents, workload, and the problems with union activities. When teachers' professional problems were analyzed seniority variable, there was little difference in scale values. While the teachers with 0-10 years experience consider decrease of professional reputation as the most important problem, the teachers with 11-45 years experience put the problems with MEB policies at the first place.

* Bu çalışma 1-3 Eylül 2016 tarihlerinde Antalya Akdeniz Üniversitesinde düzenlenen 5. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi (Uluslararası Katılımlı)'nde sözlü bildiri olarak sunulmuştur.

** Öğretmen, M.E.B., Çanakkale-Türkiye, e-posta:yaseminduyguesen@gmail.com
*** Öğretmen, M.E.B., Mersin-Türkiye, e-posta:filiz.temel30@gmail.com
**** Yrd. Doç. Dr., Ankara Üniversitesi, Eğitim Bilimleri Fakültesi, Ankara-Türkiye, e-posta:erguldemir@ankara.edu.tr

Geliş Tarihi: 03.10.2016
DOI: *10.21031/epod.297901* *Kabul Tarihi: 16.02.2017*

Keywords: Teachers' professional problems, paired comparison method, scaling

GİRİŞ

Meslek, bireyin eğitim yoluyla edinip para kazanmak için icra ettiği, sonunda hizmet türünden de olabilen bir ürün meydana getirdiği kurallı etkinlikler bütünüdür. Bu bağlamda, ürün veya hizmet meydana getirmede mesleği icra eden kişinin mesleğinden aldığı doyum önemli yer tutmaktadır. Başarı, tanınma, sosyal statü ve benzeri ihtiyaçlar ile benlik saygısının güçlenmesinde meslek icra etme kendini ifade etmenin bir yoludur denilebilir (Kuzgun, 1999). Meslek, mesleği icra edenin gereksinimlerine cevap verebildiği ölçüde hoşnutluk sebebi olabilmekte ve mesleki doyum gerçekleşebilmektedir. Mesleğin bireyin yeteneklerine uygunluğu, çalışma ortamı, maddi getirisi, ödül ve ilerleme imkânları, yüklediği sorumluluklar, çalışma süresi ve benzeri durumlar mesleki doyumu etkilemekle birlikte bu durumların bireyler üzerindeki etkileri bireylerin bunlara verdikleri önem düzeyi ile ilişkilidir. Mesleki memnuniyet ve doyum, iş verimliliğinde önemli yer tutmaktadır.

Mesleki sorun, mesleğin yürütülmesi sonucunda oluşacak ürünün veya hizmetin niteliğini olumsuz yönde etkileyen durum olarak ele alındığında, her meslek için meslek içindeki ve dışındaki sorunların tanımlanması ve bu sorunları ortadan kaldırma yöntemlerinin belirlenmesi sistemin niteliği için gereklilik arz etmektedir. Her meslek dalının kendi sisteminin gelişiminde olduğu gibi eğitim sisteminin gelişiminin devamında da doğası gereği öğretmenlik mesleği önemli bir yere sahiptir. Öğretmenlerin, eğitim sisteminin temel paydaşlardan biri olduğu düşünüldüğünde öğretmen memnuniyetinin sistemin gelişiminde ve gelişiminin devamlılığında ivme sağlayacağı açıktır. Bu bağlamda, mesleğin kendi içindeki ve dışındaki sorunların varlığı, mesleği icra edenlerin mesleki doyum ve memnuniyetini olumsuz yönde etkileyecek ve sistemin bütününün gelişimini veya gelişiminin hızını sekteye uğratacaktır.

Rudow (2002), Maslach ve Jackson (1984), araştırmaları sonucunda öğretmenlerin mesleki sorunlarının genel anlamda "eğitim sistemindeki değişiklikler, mesleki saygınlık durumu, sınıf mevcudunun fazlalığı, öğrenci davranışlarındaki olumsuzluklar, ağır iş yükü, veli ilgisizliği, meslektaş ve okul idaresinin yetersiz yardımı" olduğunu belirtmekte, Almiala (2008) ise araştırmasında söz konusu çalışmalara ek olarak nitelikli öğretmenlerin mesleklerinin ilk yıllarında istifa ettiklerini ve istifa sebebi olarak ağır iş yükünü belirttiklerini ortaya çıkarmıştır (Akt.Zarisfizadeh, 2012). Bunun yanı sıra, Bauer ve diğerleri (2007) öğretmenlik mesleğinin mesleki sorunlarının arasında eğitim sisteminin çok hızlı ve sık değiştirilmesini, kalabalık sınıfları, öğrenci davranışlarındaki olumsuzlukları, ağır iş yükünü, mesleğin statü problemini, öğretmenlerin meslektaşlarından ve idarecilerinden bekledikleri desteği görememeyi saymaktadır.

Zarisfizadeh (2012), öğretmenlerle yaptığı bir araştırmada mesleki doyumun önündeki önemli engellerin "kişisel gelişim imkânının sağlanmaması, aile ve kişisel yaşama ayrılacak vaktin azlığı, iş güvenliği ve maaşların yetersizliği" olduğunu saptamıştır. Bu sıralamada ekonomik yetersizliğin sonlarda yer aldığına, aileye ayrılacak vaktin yetersizliğinin ise ilk sıralarda yer aldığına vurgu yapılmıştır.

Türkiye'deki ilgili araştırmalar incelendiğinde, Türkiye'de öğretmenlerin mesleki ve toplumsal statüleri, güçleri, toplumsal değişmeye etkileri ve sorunlarını dönemler itibariyle ve kapsamlı olarak ortaya koyan araştırmalardan biri Akyüz (1978) tarafından yürütülmüştür. Bu araştırmada Akyüz, öğretmen sorunlarını mesleki ve meslek dışı olmak üzere iki grup olarak ortaya çıkarmış, mesleki olanları; öğretmenlerin sayısal durumu, öğretmen yetiştirilmesi sorunu, öğretmenlerin hukuki statüleri, öğretmenlerin ekonomik sorunları, öğretmenlerin örgütlenme sorunları ve öğretmenlerin mesleki yayınları başlıkları altında toplamıştır. Daha sonra yapılan araştırmalar incelendiğinde, birçoğunda Akyüz'ün bu sınıflaması doğrultusunda süreli yayınlar taranarak öğretmen sorunlarının ele alındığı görülmektedir (Güler, 1999; Aydın, 1999; Maraşlı 2007).

Ekinci (2010) ise, ilköğretim okulu müdürleri ve öğretmenlerin mesleki sorunlarını belirlemeye yönelik yürüttüğü çalışmasında, 39 okul müdürü ve 63 öğretmenin görüşlerini almıştır. İçerik analizi yöntemini kullanarak elde ettiği bulgulara göre, öğretmenlerin mesleki sorunlarını; okulların fiziki

sorunları, veli ve sosyal çevreden kaynaklı sorunlar, okul yöneticilerinden kaynaklı sorunlar, donanım yetersizliklerine dair sorunlar, eğitim sistemi ve üst yönetimlerden kaynaklı sorunlar, hizmet içi eğitim ihtiyacına dair sorunlar, denetim ve değerlendirmeye dair sorunlar, kırtasiyeciliğe dair sorunlar temaları şeklinde belirlenmiştir. Araştırma bulguları, öğretmen sorunlarına Akyüz'ün sınıflamasından farklı boyutlar eklendiğini göstermektedir. Bu sonuç mesleki sorunların dinamik bir yapıda olduğuna bir kanıt olarak görülebilir. Dolayısıyla, bu sorunların güncel olarak belirlenmesi ve önem düzeyinin ortaya çıkarılması önemli görülmektedir. Bu belirleme ve ortaya çıkarma için farklı ölçekleme teknikleri kullanılabilir.

Ölçekleme, uyarıcılar hakkında varılan yargıların üzerinde yapılan matematiksel işlemlerdir. Ölçekleme yapmanın amacı gözlemler sonucunda elde edilen verilere uygulanan istatistiksel işlemler sonucunda iyi nitelikli bir ölçek elde etmektir. Ölçeklemede kullanılan yöntemlerin "yargı yaklaşımı" ve "tepki yaklaşımı" olarak iki deneysel yaklaşım altında gruplandığı görülmektedir (Turgut ve Baykul, 1992).

Tepki yaklaşımları, denek tepkilerine dayanmaktadır. Bu yaklaşıma dayalı yöntemlerin kullanıldığı durumlarda deneklerden uyarıcıların ölçekleme boyutundaki yerini tarafsız olarak değil kendine göre belirlemesi istenmektedir. Yargı yaklaşımları ise, uyarıcı merkezli olup gözlemci yargılarına dayanmaktadır. Bu yaklaşıma dayalı yöntemlerin kullanıldığı durumlarda gözlemcilerden her bir uyarıcının yerini diğer uyarıcılara göre belirlemesi istenmektedir. Gözlemcilerin herhangi bir uyarıcı için belirttiği yargıların ortalama değeri o uyarıcının ölçek değeri olarak tayin edilmektedir. Tepki yaklaşımındaki sübjektifliğe karşın yargı yaklaşımında gözlemcilerden kendi sübjektif kararları değil, uyarıcıları diğer uyarıcılara bağlı olarak olabildiğince objektif biçimde karar vermeleri beklenir. (Turgut ve Baykul, 1992).

İkili Karşılaştırmalar Yöntemi (İKY), 1927 yılında Thurstone tarafından ortaya konan ve temeli Thurstone'nin Karşılaştırmalı Yargılar Kanunu'na dayanan bir ölçekleme yöntemidir. Yargıcı kararlarına dayalı bu yöntem, gözlemcilere uyarıcıların ikişer ikişer verilebilme imkânı olan her durumda kullanılabilmektedir. İlk defa, tutum maddelerinin ölçeklenmesi için kullanılmıştır. Daha sonraki dönemde de daha çok duyuşsal davranışların ölçeklenmesinde tercih edildiği görülmektedir (Turgut ve Baykul, 1992).

İkili Karşılaştırmalar Yöntemi'nde birey, uyarıcılar takımından ikişer ikişer uyarıcı ile karşı karşıya bırakılır. Birey, karşılaştığı her ikişer uyarıcıyı algıladığı haliyle büyüklüklerine göre sıralayarak bir yargıya varır. Bu sıralama sonucunda birey aynı zamanda uyarıcıları ölçekleme boyutunda bir nokta ile de eşleştirmiş olur (Thurstone, 1994). Thurstone, İkili Karşılaştırma Yöntemi'nin aksiyomlarını şu şekilde sıralamıştır:

Aksiyom 1 = Birey, ayırt etme süreci sonunda uyarıcıyı ölçekleme boyutunda (psikolojik boyut) bir nokta ile eşler.

Aksiyom 2 = Bireydeki anlık değişimler yüzünden, bir uyarıcıya farklı bireyler tarafından veya farklı zamanlarda aynı birey tarafından atanan farklı noktaların ölçekleme boyutundaki (psikolojik boyut) dağılımı normaldir.

Aksiyom 3 = Noktaların ölçekleme boyutundaki dağılımının ortalaması uyarıcının ölçek değeridir; standart kayması da ayırt etme dağılımının standart kaymasıdır.

Bireydeki anlık değişimler yüzünden, bir uyarıcıya farklı bireyler tarafından veya farklı zamanlarda aynı birey tarafından ölçekleme boyutunda atanan noktaların farklı olması hatayı doğurur. Bu hata, bireyin uyarıcılar arasındaki farkı ayırt etme yargısındaki hatadır. Ayırt etme yargılarındaki hata ile uyarıcının birey tarafından algılanan değerinin toplamı uyarıcının psikolojik boyutu dediğimiz ölçekleme boyutundaki değerine eşittir (Turgut ve Baykul, 1992).

İlgili araştırmalar incelendiğinde, ikili karşılaştırma yöntemi ile ölçeklemenin sınırlı sayıda çalışmaya konu olduğu görülmektedir. Öğretmen adayları üzerine yapılan çalışmalar ise sosyal aktivite tercihlerini (Göksel ve Polat, 2014), öğretim yöntem ve tekniklerini (Gelbal ve Altun, 2014), ölçme ve değerlendirme alanı yeterliliklerini (Arık ve Kutlu, 2013) belirlemeye yönelik olarak

yapılmış ve daha dar bir aralıkta kalmıştır. Alanyazında öğretmenlerin mesleki sorunlarının önem derecesini belirlemeye yönelik doğrudan bir araştırmaya rastlanmamakla birlikte dolaylı olarak öğretmen sorunlarını ortaya koyan Ekinci, Bindak ve Yıldırım'ın (2012) araştırma bulguları gösterilebilir. Bindak ve Yıldırım (2012), Ekinci'nin (2010) yapmış olduğu çalışma sonuçlarından yararlanarak, ilköğretim okulu yöneticilerinin öğretmenlerinin mesleki sorunlarına empatik yaklaşımlarını ikili karşılaştırma yöntemi kullanarak incelemişlerdir. Çalışmanın sonucunda öğretmenlerle yöneticilerin "mesleki sorunları" önem düzeylerine göre sıralama davranışlarının farklı olduğu; öğretmenlerce en önemli mesleki sorun olarak "veli ilgisizliği"nin ölçeklendiği rapor edilmiştir. Bu bağlamda, öğretmenlerin mesleki sorunlarının önem düzeyinin öğretmen yargılarına dayanarak ikili karşılaştırma yöntemi ile ölçeklenmesi araştırmanın temel çıkış noktası olmuştur.

Araştırmanın Amacı

Bu çalışmanın amacı; öğretmenlerin karşılaştıkları mesleki sorunları belirlemek ve bu sorunları temsil eden sorun alanlarının önem düzeyini öğretmenlerin kendi yargılarına dayalı olarak ikili karşılaştırma yöntemi ile ölçekleyebilmektir. Bu amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır:

1. Öğretmenlere göre "Öğretmenlik Mesleğinin Sorunları" nelerdir? Bu sorunlar tematik olarak nasıl gruplanmaktadır?
2. Öğretmenlerin mesleklerine ilişkin sorun alanlarının önem düzeyinin ölçek değerlerine göre sıralaması nasıldır?
3. Meslek kıdemine göre öğretmenlerin mesleki sorun alanlarının önem düzeyinin ölçek değerlerine göre sıralaması nasıldır?

YÖNTEM

Araştırma Modeli

Bu çalışma, tarama modelinde yürütülmüştür. Tarama modelleri çok sayıda elemanın yer aldığı bir evrende, evren ile ilgili karakteristik özellikleri ortaya koymayı amaçlamaktadır (Frankell, Wallen ve Hyun, 2011).

Çalışma Grubu

Bu çalışma, Türkiye'de 2015-2016 eğitim ve öğretim yılında M.E.B.'e bağlı devlet okullarında görev yapmakta olan 484 öğretmen üzerinde yürütülmüştür. Örneklem, olasılıklı olmayan örnekleme yöntemlerinden maksimum çeşitlilik örneklemesine göre belirlenmiştir. Bu kapsamda, farklı cinsiyet gruplarından, farklı eğitim kademelerinden ve farklı illerden öğretmenlerin örneklemde yer almasına dikkat edilmiştir. Araştırma soruları paralelinde örneklemde yer alan öğretmenlerin, cinsiyetlerine ve mesleki kıdemlerine göre dağılımı Tablo 1'de verilmiştir.

Tablo 1. Örneklemin Cinsiyet ve Mesleki Kıdemlerine göre Dağılımı

	Mesleki Kıdem			
	0-10 yıl	11-45 yıl	Toplam	Yüzde
Kadın	140	121	261	54
Erkek	81	142	223	46
Toplam	221	263	484	
Yüzde	46	54		

Tablo 1'de görüldüğü gibi, mesleki kıdem yılına göre sırasıyla 221'i 0-10 yıl, 263'ü 11-45 yıllık kıdeme sahip olmak üzere toplam 484 öğretmenin araştırmaya katılmıştır. Kıdem gruplarının oluşturulmasında, grup büyüklüklerinin yakın olmasına dikkat edilmiştir. Buna göre oluşturulan iki gruptan ilki düşük kıdem grubunu, ikincisi ise yüksek kıdem grubunu temsil etmektedir. Bunların 261'inin kadın, 223'ünün erkek olduğu görülmektedir. Örneklemde yer alan öğretmenlerden yaklaşık olarak 234'ü (%48) ilköğretim, 250'si (%52) ortaöğretim kademesinde görev yapmaktadır. Ayrıca öğretmenlerin yaklaşık olarak 170'i (%35) Mersin ilinde, 150'si (%31) Çanakkale ilinde, 70'i (%14) Şırnak ilinde ve 94'ü (%20) ise diğer illerde görev yapmaktadır.

Veri Toplama Aracı

Bu çalışmada veri toplama aracı olarak araştırmacılar tarafından geliştirilmiş olan "Öğretmenlik Mesleği Mesleki Sorunlar Formu" kullanılmıştır. Hazırlanma aşamasında; öncelikle alan yazın taranmış, öğretmenlik mesleğine yönelik sık gözlenen sorun alanları belirlenmiştir. Bunlar çoğunlukla ekonomik sorunlar, idari sorunlar, fiziksel ve alt yapı ile ilgili sorunlar, öğretmenlik mesleğinin konumu ile ilgili sorunlar, öğretmen ve velilerle ilgili sorunlar olarak ortaya çıkmaktadır. Sorunların değişim göstermiş olma olasılığı dikkate alınarak alan yazın taramasının dışında 48 öğretmene kompozisyon yazdırılmıştır. Bu kompozisyonlar, dokuman analizi ile analiz edilmiştir. Öğretmenlerin sorunları ve sorun alanlarını çok açık ve belirgin şekilde tanımladıkları görülmüştür. Uzman görüşleri de alınarak bu kapsamda 10 sorun alanı belirlenerek deneme formu oluşturulmuş ve yaklaşık 40 öğretmene deneme uygulaması yapılmıştır. Lisansüstü eğitim gören öğretmenler, araştırma görevlileri ve öğretim üyelerinden oluşan 12 uzmandan görüş alınmıştır. Uzman görüşleri arasındaki uzlaşma düzeyine yönelik uzlaşma katsayısı .90 olarak elde edilmiştir. Lawshe'ye (1975) göre bu düzey, gerek kapsam gerek içerik açısından uzmanların uzlaşma düzeyini gösteren çok yüksek bir geçerlik düzeyinin kanıtıdır. Dönütler doğrultusunda form yeniden düzenlenerek esas uygulamaya hazır hale getirilmiştir. Gerek kompozisyon yazdırma çalışmasına gerek deneme uygulamasına katılan öğretmenler esas uygulamaya dâhil edilmemiştir.

Esas form, kişisel bilgilerin sorgulandığı birinci bölüm ve mesleki sorun çiftlerinin karşılaştırıldığı ikinci bölümden oluşmaktadır. Ölçme aracının birinci bölümünde cinsiyet, mesleki kıdem, medeni hal, çocuk sayısı, eğitim durumu ve mezun olunan fakülte değişkenlerine yer verilmiş; ikinci bölümünde ise belirlenen 10 mesleki sorunun ikili olarak karşılaştırıldığı ve hangisinin diğerine göre daha öncelikli olduğunun belirtilmesinin istendiği toplam 45 mesleki sorun çifti yer almaktadır.

Ölçme aracı, baskı ortamı ve elektronik ortamın her ikisinde de hazırlanarak gönüllü katılımcılara kendi istekleri doğrultusunda elden, posta yoluyla ve eposta yoluyla verilmiştir.

Verilerin Analizi

Birinci araştırma sorusunun analizi için alan taraması yapılmış, buna ek olarak 48 öğretmene kompozisyon yazdırılmış ve bunlar üzerinde belge taraması yapılarak sorunlar 10 tema altında toplanmıştır.

İkinci ve üçüncü araştırma sorularının analizi için ikili karşılaştırma ile ölçekleme yöntemlerinden olan öncelikle V. Hal denklemi ile ölçekleme işlemi gerçekleştirilmiştir. Analizler öncesinde veriler kayıp veriler açısından kontrol edilmiş her bir ikili karşılaştırma düzeyinde %1'in altında kayıp veri olduğu görülmüştür. Veriler sınıflama ölçeğinde olduğundan dolayı medyan atama yöntemi kullanılarak eksiksiz veri seti elde edilmiş, ileri analizler bu eksiksiz veri seti üzerinde gerçekleştirilmiştir.

V. Hal denklemi ile ölçekleme gereği, ikili karşılaştırmalara ait frekans matrisi (F) oluşturulmuştur. Frekans matrisindeki her bir hücre ilgili araştırma sorusuna ait örneklem büyüklüğüne bölünerek oranlar matrisi (P) elde edilmiştir. Karşılaştırmanın yapılabilmesine imkân vermesi için oranlar matrisindeki her bir birimin z değerleri belirlenmiş ve birim normal sapmalar matrisi (Z matrisi) oluşturulmuştur. Z matrisinde her bir uyarıcı için ortalama z değerleri belirlenerek ölçek değerleri

(Sj) hesaplanmıştır. Ölçek değerlerinin yerinin tespiti için en küçük ölçek değeri eksen başlangıç noktası seçilerek ortaya çıkan son ölçek değerleri (Sc) sayı doğrusunda konumlandırılmıştır. Daha sonra ayırt etme yargıları varyanslarını eşit kabul eden Thurstone'nun V. Hal Denklemi'nin şartlarının sağlanıp sağlanmadığı ve ortalama hata miktarının manidar olup olmadığı ki-kare testi ile test edilmiştir (χ^2=127,60; sd=36, p<.05). Hesaplanan Ki-Kare değeri, kritik Ki-Kare değeri (χ^2=99,704; sd=36, p<.05) ile karşılaştırılmış (127,60>99,704; sd=36, p<.05) ve manidar bulunmuştur. Bu doğrultuda V. Hal denkleminin varsayımlarının karşılanmadığı görülerek ölçekleme işleminde Thurstone'nun III. Hal denkleminin kullanılmasına karar verilmiştir. Bunun için öncelikle ayırt etme yargılarının standart kaymaları hesaplanmıştır. Standart kaymaların kareleri alınarak varyans değerleri bulunmuştur. Varyans değerleri ikişer ikişer toplanarak varyans toplamları matrisi elde edilmiştir. Bu matristeki elemanların karekökleri alınmış ve varyans toplamları karekökü matrisi elde edilmiştir. Varyans toplamlarının karekökü matrisi ile birim normal sapmalar matrisinin esas köşegeni üzerindeki tüm değerler çarpılarak S matrisi elde edilmiştir. Daha sonra S matrisinde ölçekleme, V. Hal denkleminde bulunan Z matrisindeki gibi sütun ortalamaları alınarak ve en küçük ortalama sıfıra getirilerek tamamlanmıştır.

BULGULAR ve TARTIŞMA

Birinci Araştırma Sorusuna Ait Bulgular

Araştırmanın birinci alt amacı olarak "öğretmenlik mesleğinin sorunları nelerdir?" sorusuna yanıt aranmıştır. Öğretmenlik mesleği sorunlarının belirlenmesi amacıyla, ilgili araştırmaların bulgularının yanı sıra bu sorunların güncelliğini koruyup korumadığını anlamak ve varsa ek sorunları tespit etmek için gönüllülük esasına dayalı olarak 41 öğretmene kompozisyon yazdırılmış, 5 öğretmen ile grup görüşmesi ve grup görüşmesine katılamayan 1 öğretmen ile de yüz yüze görüşme yapılmıştır. Görüşme yapılan öğretmenlerin belirlenmesinde, branş ve öğretim kademesine göre çeşitlilik oluşturulmaya çalışılmıştır. Kompozisyon ve görüşmeler üzerinde belge taramaya dayalı analizler yapılmış, öğretmenlik mesleği sorunları 9 tema altında tanımlanmıştır. Bu temalar, temalar altında yer alan sorunlar ve bu sorunların belirtilme sıklığı Tablo 2'de verilmektedir.

Tablo 2. Öğretmenlik Mesleği Sorunlarının Tematik Dağılımı

Temalar	Sorunlar	F
Bakanlık Politikalarından Kaynaklı Sorunlar	Milli Eğitim Bakanlığı'nın aile eğitimi üzerine eğilmemesi (f=4), Zorunlu hizmete tabi okullarda ulaşım sorununa karşı ilgisiz kalınması (f=2), Eğitim politikalarının sürekli değiştirilmesi (f=14), 4+4+4 aksaklıkları (f=3), Ders kitaplarının niteliği (f=3), Ders saatlerinin yeniden düzenlenmesi gerekliliğine ilgisiz kalınması (f=3), Öğrenci değerlendirme ölçütlerine yönelik sorunlar (f=4), Hizmet içi seminerleri verenlerin o alanda uzman olmayışı (f=2), Öğretim programlarının esnek olmaması (f=3), Öğretmen görüşlerinin ciddiye alınmaması (f=2), Eğitim politikalarının siyasi temelli olması (f=1), Veli etkilemede güçsüz bırakılmak (f=3), Öğretmenin eğitim paydaşları arasında yalnız bırakılması (f=13), Öğretmenin kariyer basamaklarının olmayışı (f=4), Öğretmenevlerinde öncelik sahibi olmamak (f=2), Öğretmenlerin değerlendirilmesine dair sorunlar (f=2), Öğretim programlarının niteliği (f=6)	71
Eğitim Öğretim Süreçlerinden Kaynaklı Sorunlar	Okulların fiziki yetersizliği (f=10), Ölçme ve Değerlendirme Sınav Hizmetleri Genel Müdürlüğü'nün testlerinin yetersizliği (f=1), Veli ilgisizliği (f=11), Öğrencilerin hazırbulunuşluk düzeyindeki yetersizlikler (f=3), Kurullarda alınan kararların uygulamaya dönüştürülmemesi (f=2), Mesleki Tükenmişlik (f=2), Veli baskısı (f=1), Rehberlik dersine olan olumsuz tutum (f=1), Rehberlik servisinin amacının anlaşılmaması (f=1), Öğrenci ilgisizliği (f=7), Okullarda öğretmenler için yemekhane, kreş vs imkânlarının olmaması (f=4)	43
Yöneticilerden Kaynaklı Sorunlar	Yöneticilerin gönüllülük esaslı işleri zorunlu tutmaları (f=2), Yöneticilerin bireysel siyasi görüşlerden arınık objektif bir tutum sergilememesi (f=6), Yöneticilik niteliği (f=10), Alan dışı görevlerin yapılması (f=1), Yönetici baskısı (f=4), Yöneticilerin ders programı hazırlarken özenli olmaması (f=3), Yöneticilerin keyfi uygulamaları	32

	(f=6)	
Ekonomik Sorunlar	Maaşların yetersizliği (f=15), Ekonomik yardımların yetersizliği (6)	21
Mesleğe Gösterilen İtibar ile İlgili Sorunlar	Toplumdan saygı görülmemesi (f=17), Yöneticilerin öğretmenlik mesleğine saygı göstermemesi (f=1)	18
Hukuki Haklarından Kaynaklı Sorunlar	Mesleğin yıpranma payı göz önüne alınarak emekliliğe yansıtılmaması (f=1), Özlük haklarının yeniden düzenlenmemesi (5), Nöbet zorunluluğu (f=2)	8
Öğretmen Yetiştirme Sürecinden Kaynaklı Sorunlar	Öğretmenlerin kendilerini yetiştirmedeki yetersizliği (f=1), Öğretmen alan bilgisi eksikliği (f=2), Üniversite eğitiminin yetersizliği (f=3)	6
Evrak Yükü ile İlgili Sorunlar	Eğitim ve öğretim süreçleri ile ilgili işler dışında verilen evrak işlerinin yoğun olması (f=6)	6
Sendikal Faaliyetlerden Kaynaklı Sorunlar	Sendikal faaliyetlerde siyasi yakınlığa göre verilen ayrıcalıklar (f=5)	5

Tablo 2'de görüldüğü gibi öğretmenlerin; "öğretmenlerin mesleki sorunları" tematik olarak frekans büyüklüğü sırasıyla bakanlık politikalarından, eğitim-öğretim süreçlerinden, yöneticilerden, ekonomik haklardan, mesleki saygınlıktan, mesleğin hukuki haklarından, öğretmen yetiştirme sürecinden, evrak yükünden ve sendikal faaliyetlerden kaynaklı olarak gördükleri tespit edilmiştir.

Öğretmen yetiştirmenin, öğretmen niteliğini belirleyici bir unsur olduğu düşünüldüğünde öğretmen yetiştirme sürecinden kaynaklı sorunlar "öğretmen yetiştirme sorunları ve öğretmen niteliği sorunları" olarak ele alınmıştır.

Mesleğin özlük haklarının ekonomik şartlarla birlikteliği (ders saati karşılığı ücretler, nöbet ücretleri, ek ders karşılığı ücretlerinin hesaplanması vs) göz önüne alındığında ekonomik sorunların "özlük hakları ve ekonomik sorunlar" olarak temalaştırması yapılmıştır.

Mesleki sorunlar arasında yüksek frekansa sahip olan eğitim ve öğretim süreçlerinden kaynaklı sorunların kendi içindeki dağılımı dikkate alındığında bu temanın ayrıştırılıp "fiziki koşullar ve alt yapı yetersizliği sorunları (f=10), veli ile ilgili sorunlar (f=12), öğrenci ile ilgili sorunlar (f=7)" olarak ele alınmıştır.

Öğretmenlerin yöneticilerden kaynaklı sorunlar olarak belirttikleri sorunlar incelendiğinde "yönetici" kavramı ile sadece "okul idaresini" kastettikleri il-ilçe yönetimlerini kastetmedikleri belirlenmiş olup bu sebeple öğretmenlerin kullandıkları "yönetici" kavramı yerine "okul idaresi" kabul edilmiş ve mesleki sorun olarak "okuldaki idari sorunlar" kabul edilmiştir.

Öğretmenlerin beyan ettikleri mesleki sorunlar arasında yer alan evrak yükü ile ilgili sorunlar incelendiğinde Eğitim ve öğretim süreçleri ile ilgili işler dışında verilen evrak işlerinin yoğun olmasının vurgu aldığı görülmektedir. Öğretmenlerin ifadelerinden eğitim ve öğretim işlerinin dışındaki iş ve işlemleri iş yükü olarak gördükleri dikkat çekmektedir. Bu sebeple Tablo 2'de belirtilmiş olan "evrak yükü ile ilgili sorunlar" teması, mesleki sorun bağlamında "iş yükünün fazlalığı" olarak ele alınmıştır. Tüm bunlar göz önüne alındığında, araştırma kapsamında yanıt aranan öğretmenlik mesleğine ait sorunlar araştırma alt problemlerine yanıt aranmasında kolaylık sağlanması bakımından kodlarla da ifade edilerek belirlenerek Tablo 3'de verilmiştir.

Tablo 3'de görüldüğü gibi öğretmenlik mesleğine ait sorunlar on başlık altında toplanmıştır. Tablo 2'de verilmiş olan tematik dağılım kavram kullanımları çerçevesinde Tablo 3'deki verilmiş olan sorunlarla ifade edilmiştir.

Tablo 3. Öğretmenlik Mesleği Sorun Alanları ve Kodları

Mesleki Sorun	Mesleki Sorun Kodu
Öğretmen yetiştirme sorunları ve öğretmen niteliği	A
Özlük hakları ve ekonomik sorunlar	B
Mesleki saygının azalması	C
M.E.B. Politikaları ile ilgili sorunlar	D
Sendikal Faaliyetlerle ilgili sorunlar	E
İş yükünün fazlalığı	F
Okuldaki idari sorunlar	G
Fiziki koşullar ve alt yapı yetersizliği	H
Veli ile ilgili yaşanan sorunlar	I
Öğrenci ile ilgili sorunlar	J

İkinci Araştırma Sorusuna Ait Bulgular

Öğretmenlere yazdırılan kompozisyonlar dokuman analizi ile incelenmiş ve on mesleki sorun alanı belirlenmiştir. Her öğretmenden belirlenen on sorun alanını ikili olarak karşılaştırmaları istenmiştir. Bu karşılaştırma sonrası her bir uyarıcıya ilişkin frekans değerleri belirlenmiştir. Belirlenen frekans değerleri Tablo 3'de belirtilmiş olan [A-J] kodlama sistemine göre Tablo 4'de verilmiştir.

Tablo 4. Öğretmenlik Mesleği Mesleki Sorun Alanlarının Frekans Matrisi

	A	B	C	D	E	F	G	H	I	J
A		217	365	384	176	223	246	328	263	294
B	267		312	328	126	158	195	266	165	242
C	119	172		247	102	128	166	201	129	179
D	100	156	237		76	116	129	170	117	163
E	308	358	382	408		298	328	345	295	332
F	261	326	356	368	186		266	305	257	299
G	238	289	318	355	156	218		284	230	274
H	156	218	283	314	139	179	200		194	238
I	221	319	355	367	189	227	254	290		285
J	190	242	305	321	152	185	210	246	199	
Toplam	1860	2297	2913	3092	1302	1732	1994	2435	1849	2306

Tablo 4'deki frekans matrisi, satırdaki uyarıcının sütundaki uyarıcıya göre tercih edilme sayısını göstermektedir. Örneğin tabloya göre ikinci satır birinci sütunda yer alan 267 öğretmen A sorun alanını B sorun alanına tercih ederken, birinci satır ikinci sütunda yer alan 217 öğretmen B sorun alanını A sorun alanına tercih etmiştir. Bir diğer anlamla, B sorun alanının A sorun alanına göre daha önemli olduğunu belirtmiştir. Aynı sorunun tercih edilmesi gibi bir durum söz konusu olmadığından köşegen üzerinde herhangi bir değer bulunmamaktadır.

Frekans matrisi ortaya çıkarıldıktan sonra frekans matrisinin her bir hücresindeki değer, ikili karşılaştırmayı yapan toplam kişi sayısı olan 484'e bölünerek oranlar matrisi bulunmuştur. Oranlar matrisine ilişkin değerler Tablo 5'de gösterilmiştir.

Tablo 5. Oranlar Matrisi

	A	B	C	D	E	F	G	H	I	J
A		0,448	0,754	0,793	0,364	0,461	0,508	0,678	0,543	0,607
B	0,552		0,645	0,678	0,26	0,326	0,403	0,55	0,341	0,5
C	0,246	0,355		0,51	0,211	0,264	0,343	0,415	0,267	0,37
D	0,207	0,322	0,49		0,157	0,24	0,267	0,351	0,242	0,337
E	0,636	0,74	0,789	0,843		0,616	0,678	0,713	0,61	0,686
F	0,539	0,674	0,736	0,76	0,384		0,55	0,63	0,531	0,618
G	0,492	0,597	0,657	0,733	0,322	0,45		0,587	0,475	0,566
H	0,322	0,45	0,585	0,649	0,287	0,37	0,413		0,401	0,492
I	0,457	0,659	0,733	0,758	0,39	0,469	0,525	0,599		0,589
J	0,393	0,5	0,63	0,663	0,314	0,382	0,434	0,508	0,411	

Tablo 5 incelendiğinde köşegene göre simetrik olan her bir oran çiftinin toplamının 1'i verdiği görülmektedir. Oranlar matrisinden ortalaması 0, standart sapması 1 olan birim normal sapmalar matrisine geçilerek z değerleri elde edilmiştir. Elde edilen z değerleri Tablo 6'da gösterilmiştir.

Tablo 6. Birim Normal Sapmalar Matrisi

	A	B	C	D	E	F	G	H	I	J
A		-0,131	0,687	0,817	-0,348	-0,098	0,02	0,462	0,108	0,272
B	0,131		0,372	0,462	-0,643	-0,451	-0,246	0,126	-0,41	0
C	-0,687	-0,372		0,025	-0,803	-0,631	-0,404	-0,215	-0,622	-0,332
D	-0,817	-0,462	-0,025		-1,007	-0,706	-0,622	-0,383	-0,7	-0,421
E	0,348	0,643	0,803	1,007		0,295	0,462	0,562	0,279	0,485
F	0,098	0,451	0,631	0,706	-0,295		0,126	0,332	0,078	0,3
G	-0,02	0,246	0,404	0,622	-0,462	-0,126		0,22	-0,063	0,166
H	-0,462	-0,126	0,215	0,383	-0,562	-0,332	-0,22		-0,251	-0,02
I	-0,108	0,41	0,622	0,7	-0,279	-0,078	0,063	0,251		0,225
J	-0,272	0	0,332	0,421	-0,485	-0,3	-0,166	0,02	-0,225	

Tablo 6 incelendiğinde esas köşegene göre birbirinin simetriği olan değerlerin işaret olarak birbirinin tersi olduğu görülmektedir. Birim normal sapmalar matrisi üzerinde gerekli işlemler yapılarak σi ve bunların karesi alınarak varyans değerleri bulunmuştur. Varyans değerleri ikişer ikişer toplanarak esas köşegenin üstünde yer alan hücrelere yazılmıştır. Bu şekilde elde edilen varyans toplamları matrisi Tablo 7'de gösterilmiştir.

Tablo 7. Varyans Toplamları Matrisi

	A (0,432)	B (0,507)	C (1,408)	D (0,835)	E (1,373)	F (1,097)	G (1,115)	H (1,264)	I (0,869)	J (1,437)
A (0,432)		0,939	1,84	1,267	1,805	1,529	1,547	1,696	1,301	1,869
B (0,507)			1,915	1,342	1,88	1,604	1,622	1,771	1,376	1,944
C (1,408)				2,243	2,781	2,505	2,523	2,672	2,277	2,845
D (0,835)					2,208	1,932	1,95	2,099	1,704	2,272
E (1,373)						2,47	2,488	2,637	2,242	2,81
F (1,097)							2,212	2,361	1,966	2,534
G (1,115)								2,379	1,984	2,552
H (1,264)									2,133	2,701
I (0,869)										2,306
J (1,437)										

Varyans toplamları matrisindeki değerlerin tümünün karekökü alınarak varyans toplamlarının karekökü matrisi elde edilmiştir. Varyans toplamlarının karekökü matrisi Tablo 8'de verilmiştir.

Tablo 8. Varyans Toplamlarının Karekökü

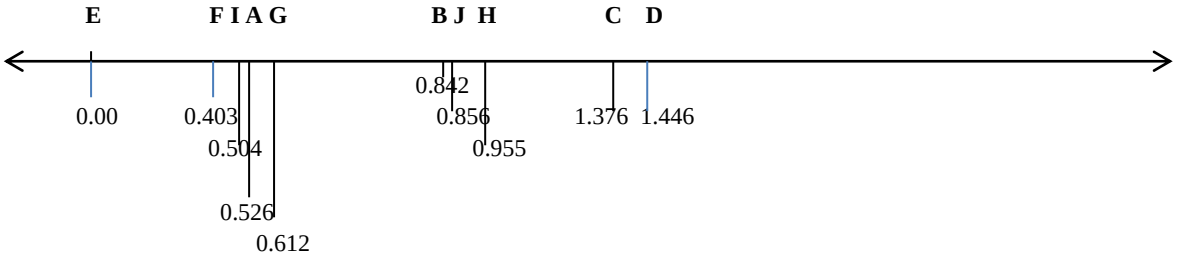
	A	B	C	D	E	F	G	H	I	J
A		0,969	1,356	1,126	1,344	1,237	1,244	1,302	1,141	1,367
B			1,384	1,158	1,371	1,266	1,274	1,331	1,173	1,394
C				1,498	1,668	1,583	1,588	1,635	1,509	1,687
D					1,486	1,390	1,396	1,449	1,305	1,507
E						1,572	1,577	1,624	1,497	1,676
F							1,487	1,537	1,402	1,592
G								1,542	1,409	1,597
H									1,460	1,643
I										1,519
J										

Tablo 8'deki varyans toplamlarının karekökü matrisi ile birim normal sapmalar matrisinin esas köşegeni üstünde kalan tüm değerler çarpılarak S matrisi elde edilmiştir. Elde edilen S matrisi Tablo 9'da verilmiştir.

Tablo 9. Öğretmenlerin Sorun Alanlarına İlişkin S Matrisi

	A	B	C	D	E	F	G	H	I	J
A		-0,052	0,898	0,703	-0,451	-0,118	0,024	0,58	0,135	0,492
B	0,052		0,494	0,413	-0,847	-0,548	-0,301	0,161	-0,521	0
C	-0,898	-0,494		0,032	-1,282	-0,955	-0,614	-0,336	-0,97	-0,675
D	-0,703	-0,413	-0,032		-1,266	-0,812	-0,72	-0,465	-0,633	0,354
E	0,451	0,847	1,282	1,266		0,443	0,697	0,873	0,432	0,983
F	0,118	0,548	0,955	0,812	-0,443		0,179	0,488	0,114	0,588
G	-0,024	0,301	0,614	0,72	-0,697	-0,179		0,324	-0,093	0,326
H	-0,58	-0,161	0,336	0,465	-0,873	-0,488	-0,324		-0,38	-0,04
I	-0,135	0,521	0,97	0,633	-0,432	-0,114	0,093	0,38		0,449
J	-0,492	0	0,675	-0,354	-0,983	-0,588	-0,326	0,04	-0,449	
Toplam	-2,211	1,097	6,192	4,69	-7,274	-3,359	-1,292	2,045	-2,365	2,477
SJ	-0,221	0,11	0,619	0,469	-0,727	-0,336	-0,129	0,205	-0,237	0,248
Sc	0,506	0,837	1,346	1,196	0	0,391	0,598	0,932	0,49	0,975

Tablo 9 incelendiğinde köşegene göre elemanların birbirinin simetriği olan değerlerin işaret olarak birbirinin tersi olduğu görülmektedir. Her bir sütuna ait toplam değerlerin gösterebilmek için matrisin sonuna bir satır eklenerek sütun toplamları alınmıştır. Daha sonra sütun toplamları uyarıcı sayısı olan 10 sayısına bölünmüş ve Sj değerleri elde edilmiştir. Bir sonraki adımda ise eksenin başlangıç noktasını Sj satırındaki en küçük değer olan -0,727 sayısına kaydırmak için her bir ölçek değerine bu değerin mutlak değeri olan 0,727 sayısı eklenerek her bir sorunun ölçek değeri belirlenmiş ve Sc satırında belirtilmiştir. Her bir mesleki sorunun ölçek değeri Şekil 1'de sayı doğrusu üzerinde sıralanmıştır.



Şekil 1. Öğretmenlik Mesleği Mesleki Sorunlarının Ölçek Değerleri

Şekil 1'deki ölçek değerlerine göre öğretmenlerin sorun alanlarının önem sırası belirlenmiş ve Tablo 10' da kod açılımlarıyla belirtilmiştir.

Tablo 10. Öğretmenlerin Mesleki Sorun Alanlarının Ölçek Değerleri ve Uyarıcı Sıraları

Öğretmenlerin Mesleki Sorunları	Ölçek Değerleri	Uyarıcı Sıraları
Öğretmen Yetiştirme Sorunları ve Öğretmen Niteliği (A)	0,526	7
Özlük Hakları ve Ekonomik Sorunlar (B)	0,842	5
Mesleki Saygınlığın Azalması (C)	1,376	2
MEB Politikaları ile ilgili Sorunlar (D)	1,446	1
Sendikal Faaliyetlerle İlgili Sorunlar €	0	10
İş Yükü Fazlalığı (F)	0,403	9
Okuldaki İdari Sorunlar (G)	0,612	6
Fiziki Koşullar ve Alt yapı Yetersizliği (H)	0,955	3
Veli ile İlgili Yaşanan Sorunlar (I)	0,504	8
Öğrenci ile İlgili Yaşanan Sorunlar (J)	0,856	4

Tablo 10'da görüldüğü gibi öğretmenlerin mesleki açıdan en önemli gördükleri sorun MEB politikalarıdır. Bu sorunu sırası ile mesleki saygınlıklarının azalması, fiziki koşullar ve alt yapı yetersizliği, öğrenci ile ilgili yaşanan sorunlar, özlük hakları ve ekonomik sorunlar, okuldaki idari sorunlar, öğretmen yetiştirme ve öğretmen niteliği, veli ile ilgili yaşanan sorunlar, iş yükü fazlalığı ve sendikal faaliyetlerle ilgili sorunlar izlemektedir.

Üçüncü Araştırma Sorusuna Ait Bulgular

Bu başlıkta 3. Araştırma sorusuna yanıt aranmıştır. Bunun için öğretmenler kıdemlerine göre iki gruba ayrılmış; 10 yıllık kıdeme sahip öğretmenler 1. grup, 11-45 yıllık kıdeme sahip öğretmenler 2. Grup olarak ele alınmıştır. Bu iki gruba göre öğretmenlerin mesleki sorunları ayrı ayrı ölçeklendiğinde elde edilen ölçek değerleri ve uyarıcı sıraları Tablo 11'de gösterilmiştir.

Tablo 11. Kıdem Gruplarına göre Öğretmenlerin Mesleki Sorunları

Öğretmenlerin Mesleki Sorunları	0-10 yıl kıdem		11-45 yıl kıdem	
	Ölçek Değerleri	Uyarıcı Sıraları	Ölçek Değerleri	Uyarıcı Sıraları
Öğretmen Yetiştirme Sorunları ve Öğretmen Niteliği	0,676	8	0,447	7
Özlük Hakları ve Ekonomik Sorunlar	0,874	5	0,817	4
Mesleki Saygınlığın Azalması	1,546	1	1,126	2
MEB Politikaları ile ilgili Sorunlar	1,54	2	1,313	1
Sendikal Faaliyetlerle İlgili Sorunlar	0	10	0	10
İş Yükü Fazlalığı	0,544	9	0,364	8
Okuldaki İdari Sorunlar	0,815	6	0,491	6
Fiziki Koşullar ve Alt yapı Yetersizliği	1,065	4	0,83	3
Veli ile İlgili Yaşanan Sorunlar	0,8	7	0,306	9
Öğrenci ile İlgili Yaşanan Sorunlar	1,232	3	0,573	5

Tablo 11 incelendiğinde, kıdemlere göre öğretmenlerin mesleki sorunlarının önem derecelerinin çok küçük farklarla değiştiği görülmektedir. 0-10 yıl arası görev yapan öğretmenler için en önemli sorun mesleki saygınlığın azalması iken 11-45 yıl arası görev yapan öğretmenler için en önemli sorunu MEB politikalarıdır. 0-10 yıllık öğretmenler için ikinci sorun olarak MEB politikaları görülürken bunu öğrenci ile ilgili yaşanan sorunlar, fiziki koşullar ve alt yapı yetersizliği, özlük hakları ve ekonomik sorunlar, okuldaki idari sorunlar, veli ile ilgili yaşanan sorunlar, öğretmen yetiştirme ve öğretmen niteliği ile ilgili sorunlar, iş yükü fazlalığı ve sendikal faaliyetlerle ilgili sorunlar takip etmektedir. 11-45 yıllık öğretmenler için ise ikinci sorun olarak mesleki saygınlığın azalması olarak görülürken bunu fiziki koşullar ve alt yapı yetersizliği, özlük hakları ve ekonomik sorunlar, öğrenci ile ilgili yaşanan sorunlar, okuldaki idari sorunlar, yetiştirme ve öğretmen niteliği ile ilgili sorunlar, iş yükü fazlalığı, veli ile ilgili yaşanan sorunlar ve sendikal faaliyetlerle ilgili sorunlar takip etmektedir.

Ekinci'nin (2010) çalışmasında öğretmen sorunlarının "okulların fiziki sorunları, veli ve sosyal çevreden kaynaklı sorunlar, okul yöneticilerinden kaynaklı sorunlar, donanım yetersizliklerine dair sorunlar, eğitim sistemi ve üst yönetimlerden kaynaklı sorunlar" bağlamında elde ettiği bulgularla bu çalışmanın birinci araştırma sorusuna dair elde edilen bulguların benzer olduğu görülmüştür. Akyüz'ün (1978) belirlediği öğretmen sorunlarından "öğretmen yetiştirilmesi sorunu, öğretmenlerin hukuki statüleri, öğretmenlerin ekonomik sorunları, öğretmenlerin örgütlenme sorunları" ile paralellik göstermekte olup, mesleki yayınlar sorunlarında farklılaşma gösterdiği görülmüştür. Bu farklılıkların mesleki sorunlarının dinamik yapıda olmasından kaynaklandığı zaman içinde alınan kararlar doğrultusunda yapılan değişiklikler ile ilişkili olduğu düşünülmektedir.

Öğretmenlik mesleğinin en önemli sorunların mesleki saygınlığın azalması ve MEB politikalarının seçilmesi eğitim sistemindeki istikrarsız değişimlerden ve öğretmenlerin toplumun gözündeki değerinin azalmasından rahatsız olduklarının göstergesidir. Bu bulgunun Rudow (2002), Maslach ve Jackson'ın (1984) araştırmalarının bulgularıyla (Akt: Bauer vd, 2007) (eğitim sistemindeki değişiklikler, mesleki saygınlık durumu, sınıf mevcudunun fazlalığı, …) paralel olduğu görülmüştür.

Almiala'nın (2008) araştırma bulguları, nitelikli öğretmenlerin mesleklerinin ilk yıllarında istifa ettiklerini ve istifa sebebi olarak ağır iş yükünü belirttiklerini ortaya çıkarmıştır (Akt. Zarisfizadeh, 2012). Türkiye'de M.E.B.'e bağlı okullarda görev yapmakta olan öğretmenlerin sorunlarının belirlendiği bu çalışmada ise ağır iş yükü olarak ifade edilebilen iş yükü fazlalığının on mesleki sorun içinde dokuzuncu sırayı aldığı tespit edilmiştir. Almiala'nın (2008) araştırmasındaki bulgularla bu çalışmanın bulguları arasındaki farkın devletler arasındaki devlet güvencesi ve özel sektör bağlamından geldiği düşünülmektedir.

SONUÇ ve ÖNERİLER

Bu çalışmada öğretmenlerin, meslek hayatlarında karşılaştıkları sorunların ve buna göre belirlenen sorun alanlarının önem düzeyi, kendi yargılarına dayalı olarak belirlenmiştir. Çalışmanın birinci araştırma sorusu kapsamında elde edilen bulgulara göre öğretmenlerin mesleki sorun alanları mesleki saygınlığın azalması, MEB politikaları ile ilgili sorunlar, öğrenci ile yaşanan sorunlar, fiziki koşullar ve alt yapı yetersizliği, özlük hakları ve ekonomik sorunlar, okuldaki İdari sorunlar, öğretmen yetiştirme sorunları ve öğretmen niteliği, veli ile ilgili yaşanan sorunlar, iş yükü fazlalığı, sendikal faaliyetlerle ilgili sorunlar olmak üzere 10 tema altında toplanmıştır.

Çalışmanın ikinci araştırma sorusu kapsamında elde edilen bulgular sonucunda öğretmenlere göre belirlenen sorun alanları içerisinde ilk iki sırayı mesleki saygınlığın azalması ve MEB politikaları ile ilgili sorunların aldığı bulunmuştur. Bunları sırası ile öğrenci ile yaşanan sorunlar, fiziki koşullar ve alt yapı yetersizliği, özlük hakları ve ekonomik sorunlar, okuldaki idari sorunlar, öğretmen yetiştirme sorunları ve öğretmen niteliği, veli ile ilgili yaşanan sorunlar, iş yükü fazlalığı, sendikal faaliyetlerle ilgili sorunlar takip etmiştir.

Çalışmanın üçüncü araştırma sorusu kapsamında elde edilen bulgular sonucunda 0-10 yıllık öğretmenler için sorun alanlarında ilk iki sırayı mesleki saygınlığın azalması ve MEB politikaları almış; daha kıdemli öğretmenler ise MEB politikalarını daha önemli sorun olarak görmüştür. Her iki grup için sendikal faaliyetler en son sırada yer alarak diğerlerine göre daha az önemli olan bir sorun alanı olarak görülmüştür.

Akyüz'ün (1978) bulguları arasında olan mesleki yayınlarla ilgili sorunların bu araştırmanın bulgularından olmaması araştırmanın güncelliğine ve eğitim sistemindeki değişimin getirilerine bağlanmaktadır. Ancak bu savın geçerliğinin araştırılması önerilmekte ve mesleki sorunların yıllar bazında değişiminin incelenmesi ve izlenmesi önerilmektedir.

Almiala'nın (2008) araştırması (Akt. Zarisfizadeh, 2012) ile karşılaştırma yapabilmek için örneklemlerin birbirine uygun olarak seçilerek Türkiye'de M.E.B.'e bağlı okullarda 4/A kadrosunda görev yapmakta olan öğretmenlerin mesleklerinin ilk yıllarında istifa etme durumlarının belirlenmesi önerilmektedir.

Bu araştırma sonucunda belirlenen sorunlar ve sorun alanları üzerine nitel bir araştırma yapılarak çözüm önerilerinin belirlenmesi, öğretmenlik mesleğinin zaman içinde değişkenlik gösterme sebeplerinin ortaya çıkarılması, farklı ölçekleme teknikleri kullanılarak öğretmenlik mesleği sorunlarının önem düzeylerinin karşılaştırılmasının yapılması, öğretmenlik mesleği sorunlarının farklı değişkenler açısından önem düzeylerinin belirlenmesi önerilmektedir.

Yapılacak yeni araştırmaların öğretmenlerin mesleki sorunlarının ve sorun alanlarının derinlemesine bir bakış geliştirilmesine yardımcı olacağı düşünülmektedir. Çünkü eğitim sisteminin niteliği sistemin içindeki paydaşların mutluluğuna bağlıdır. Bireylerin mutluluğu ise karşılaşılan sorunların ortadan kaldırılmasına bağlıdır. Bu noktada, karşılaşılan sorunlara çözüm önerileri getirmek; o sorunların varlığının tespiti ve işe vuruk tanımlanması ile mümkündür. Bu sebeple öğretmenlik mesleğinin sorunlarına çözüm üretilmek isteniyor ise bu sorunların önem düzeyleri ile birlikte belirlemek ile başlamak daha doğrudur.

KAYNAKÇA

Akyüz, Y. (1978). *Türkiye'de öğretmenlerin toplumsal değişmedeki etkileri (1848-1940).* Ankara: Doğan Basımevi.

Altun, A. ve Gelbal, S. (2014). Öğretmenlerin kullandıkları ölçme ve değerlendirme yöntem veya araçlarının ikili karşılaştırma yöntemiyle belirlenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 1-11.

Arık, R. S. ve Kutlu, Ö. (2013). Öğretmenlerin ölçme ve değerlendirme alanı yeterliklerinin yargıcı kararlarına dayalı ölçeklenmesi. *Eğitim Bilimleri Araştırmaları Dergisi, 3*(2), 163-196.

Aydın, R. (1999). *Türk basınında öğretmen sorunları (1940-1955 yılları arası).* Yayınlanmamış Yüksek Lisans Tezi. Ankara Üniversitesi Eğitim Bilimleri Enstitüsü. Ankara.

Bauer, J., Unterbrink, T., Hack, A., Pfeifer, R., Buhl-Grieshaber, V., Müller, U., Wesche, H., Frommhold, M., Seibt, R., Scheuch, K., & Wirsching, M. (2007). Working conditions, adverse events and mental health problems in a sample of 949 German teachers. *International Archives of Occupational and Environmental Health, 80*(5), 442-449. Retrieved from http://link.springer.com/article/10.1007/s00420-007-0170-7#page-2

Ekinci, A. (2010). İlköğretim okullarında çalışan müdür ve öğretmenlerin mesleki sorunlarına ilişkin görüşleri. *İlköğretim Online, 9*(2), 734-748.

Ekinci, A., Bindak, R. ve Yıldırım, M. C. (2012). İlköğretim okulu yöneticilerinin öğretmenlerin mesleki sorunlarına empatik yaklaşımlarının ikili karşılaştırmalar metodu ile incelenmesi. *Gaziantep Üniversitesi Sosyal Bilimler Dergisi, 3*, 759-776.

Güler, A. V. (1999). Cumhuriyet dönemi bazı süreli yayınlara yansıyan öğretmen sorunları (1929-1961). *Yayınlanmamış Yüksek Lisans Tezi.* Ankara Üniversitesi Sosyal Bilimleri Enstitüsü. Ankara.

Kuzgun, Y., Sevim S. A. ve Hamamcı, Z. (1999). Mesleki doyum ölçeğinin geliştirilmesi. *Türk Psikolojik Danışma ve Rehberlik Dergisi, 2*(11),14-18

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Maraşlı, Ş. (2007). Türkiye'de eğitimle ilgili süreli yayınlara yansıyan öğretmen sorunları (1970-2000). *Yayınlanmamış Yüksek Lisans Tezi.* Ankara Üniversitesi Eğitim Bilimleri Enstitüsü. Ankara.

Polat, B. ve Göksel, H.Ç. (2014). Öğretmen adaylarının sosyal aktivite tercihlerinin ikili karşılaştırma yöntemiyle belirlenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 5*(1), 88-100.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review, 101,* 266-270.

Turgut, M. F. ve Baykul, Y. (1992). *Ölçekleme teknikleri.* Ankara: ÖSYM Yayınları.

Zarisfizadeh, S. (2012). Job satisfaction factors among english language teachers in Malaysia. *International Journal of Applied Linguistics and Englisg Literature, 1*(4), 30-36. Retrieved from http://www.journals.aiac.org.au/index.php/IJALEL/article/view/741/671

EXTENDED ABSTRACT

Introduction

Profession problem is a situation which negatively affects the quality of the product or service of the profession. Thus, it is highly important for the system's well-being to identify the problems from inside and outside of the profession for each occupational group and to detect the methods in order to remove them. It is clear that teachers' satisfaction will provide positive moves both for the development and the sustainability of the system as teachers are the shareholders of the education system. Considered in this context, the presence of inner or outer problems does negatively affect the professional satisfaction and content while it also will interrupt the development or growth rate of the system.

In the related literature, there are many researches on teachers' profession problems (Almiala, 2008; Maslach and Jackson, 1984; Rudow, 2002; Zarisfizadeh, 2012). Also, this concept has been studied in Turkey (Akyüz, 1978; Aydın, 1999; Ekinci, 2010; Güler, 1999; Maraşlı, 2007). All these studies' findings show that professional problems have a dynamic nature. That's why, it is important to update them and identify the relative importance. Various scaling methods can be used for this aim.

Scaling is a mathematical process for the judgement regarding the stimulus. The aim of scaling process is to develop a well-qualified scale as a result of the statistics used to analyze the data collected. There are two experimental methods for scaling named as 'judgmental approach' and 'reaction approach' (Turgut ve Baykul, 1992).

When the related literature is studied, it is revealed that pairwise comparison method has been used for a very limited number of researches. There has been found no direct research regarding the importance ranking of teachers' professional problems. Yet, Ekinci, Bindak and Yıldırım (2012) used the findings of Ekinci's research (2010) and investigated the primary school managers' attitude towards teachers' professional problems via pairwise comparison method. Study showed that there was a difference between the importance ranking of professional problems between teachers and school managers. It was found out that teachers and managers' ranking behaviors of professional problems showed difference. Teachers focused on the parents' indifference as a major professional problem.

The aim of this study is to scale the importance level of professional problems that the teachers encounter using the pairwise comparison method. The questions to be answered as follows:

1. What are the problems of teaching profession according to the teachers?
2. How are the rankings of the importance levels of teachers' professional problems according to the scale values?
3. How are the rankings of the importance level of teachers' professional problems on the basis of teachers' seniority according to the scale values?

Method

This research is designed as a survey model. Study group consisted of 484 secondary school teachers who worked for public school in 2015-2016 academic year in Turkey. Sample was selected using the maximum variety sampling out of nonprobability sampling methods.

Teaching Profession Problems Form was developed and used to gather data. The first part of the form includes questions regarding personal information while the second part consists of the 45 pair of profession problems which are to be compared pairwise.

In order to answer the first research question, teachers were asked to write compositions. Those compositions were analyzed via documents analysis and grouped under ten themes. To answer the second and the third research questions, pairwise comparison scaling was used. Data set was controlled for missing values. As the assumptions of V. equation of state were not met (χ^2=169,126, df=36 and $p<0.05$), it was decided to use III. Equation of state.

Findings

As an answer for the first research question, teaching profession problems were seen to get factored under ten themes. Those are teacher training issues-teacher quality (A), personal benefits and financial issues (B), prestige loss (C), ministry of national education policies (D), organizational events (E), overloaded work (F), managerial problems at school (G), physical conditions (H), parent issues (I) and student issues (J).

As an answer to second research question, the scale values and the ranks of profession problems were calculated. According to this, teachers consider the ministry's policies and decrease of professional reputation as their major professional problems. Hard infrastructure and conditions of the schools, problems with students, personal benefits, and economic conditions are respectively important.

In order to answer the third research question, teachers are divided into two groups according to their seniority. Scale values and stimulus ranking obtained from these groups were calculated. According to findings, there are differences between teachers in terms of their seniority. It can be concluded that these differences are not much but considerable. Organizational events are considered as the least important problem in both seniority levels. Overloaded work, parent issues, and teachers' quality are seen relatively less important. Prestige loss and Ministry of National Education policies are considered as the most important problems followed with student issues, physical conditions, personal benefits and financial problems, and managerial problems at schools.

Results and Discussion

It was found that the findings of this research are parallel with the findings of Ekinci's study. The profession problems that Akyüz (1978) identified as teachers' qualification issue, legal issue, financial issue, and organizational issue are also parallel with these research findings. There are differences between findings because of the nature of the teaching profession.

Major problems identified by teachers are prestige loss and ministry of national education policies. These major problems are the signs that teachers feel uncomfortable about the unstable education policies and their loss of prestige in public. This finding is parallel with the findings of Rudow (2002) and Maslach and Jackson (1984) studies (cited from Bauer vd, 2007).

It can be suggested in the light of this study findings that effort to fix the educational policies and regain of professional prestige should be organized. Other shareholders of education system should be taken into consideration for further research. It also can be suggested to conduct such researches with qualitative data.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);63-78

Puanlayıcılar Arası Güvenirlik Belirleme Tekniklerinin Karşılaştırılması*

The Comparison of Interrater Reliability Estimating Techniques

Özge BIKMAZ BİLGEN ** Nuri DOĞAN ***

Öz

Bu araştırmada dereceli puanlama anahtarı türü ve puanlayıcı sayısı değişiminin, puanlayıcı güvenirliğini belirlemede kullanılan tekniklerden elde edilen sonuçlar üzerindeki etkisi incelenmiştir. Araştırmanın çalışma grubu, 50 öğrenci ve puanlama yapan 10 öğretmenden oluşmaktadır. Betimsel nitelik taşıyan araştırmada puanlayıcı güvenirliğini belirlemede Kappa istatistik tekniği, log linear analiz tekniği ve Krippendorff alfa tekniği kullanılmıştır. Puanlayıcı sayısı değişiminin puanlayıcı güvenirliğine etkisini incelemek adına belirtilen üç teknik kullanılarak iki, beş ve on puanlayıcı arasındaki uyum düzeyleri hesaplanmıştır. Araştırmada üç teknikten elde edilen analiz sonuçlarında, analitik puanlama anahtarı kullanımıyla elde edilen puanlarda, puanlayıcı sayısı artışının güvenirlik düzeyini düşürdüğü tespit edilmiştir. Üç teknikle yapılan analizlerde, en yüksek güvenirlik değerleri iki puanlayıcı kullanıldığında elde edilmiş, puanlayıcı sayısı artırıldıkça güvenirliğin düştüğü saptanmıştır. Analitik puanlama anahtarını oluşturan kategoriler incelendiğinde kategoriler arasında objektiflik düzeyine dayalı olarak, puanlayıcıların uyum düzeylerinde değişkenlik olduğu saptanmıştır. Araştırmanın sonucunda, kullanılan tekniklerden Kappa tekniği ve Krippendorff alfa tekniğinin paralel sonuçlar verdiği görülmüştür. Bununla birlikte Krippendorff alfa tekniğinin puanlayıcı sayısı değişiminden Kappa tekniğine göre daha az etkilendiği belirlenmiştir. Log-linear analiz tekniğinin ise değişkenler arasındaki etkileşimleri ve uyumsuzluk kaynağını gösteren daha kapsamlı ve geniş bilgi sağladığı tespit edilmiştir. Sonuç olarak, daha detaylı ölçme sonuçları elde edilmek istendiğinde alt kategorilerden oluşan analitik puanlama anahtarı kullanılarak toplanan puanların, kategorik veri analizi için uygun olan log-linear analiz tekniğinin; daha genel ölçme sonuçlarına ulaşılmak istendiğinde ise bütünsel puanlama anahtarı ile elde edilen puanların Krippendorff alfa tekniğinin kullanılmasının uygun olduğu düşünülmektedir.

Anahtar Kelimeler: Kappa istatistiği, log-linear analiz tekniği, Krippendorff alfa

Abstract

The aim of this study is to analyse the effects of the number of raters and the types of rubric on the results obtained by the techniques used to estimate the interrater reliability. The research group consists of 50 students and 10 teachers who rated. As a descriptive study, in this paper the Kappa statistical technique, the log linear analysis technique, and the Krippendorff alpha technique were used to determine the rater reliability. In order to investigate the effects of the number of raters on the interrater reliability, the level of agreement between 2, 5, and 10 raters was calculated by using those three techniques. The findings obtained from the three techniques demonstrated that the use of analytic rubric provided much more reliable ratings than holistic rubric. Moreover, it was also found based on the analysis results obtained through all three techniques that maximum reliability values were obtained by using two raters, reliability values decreased with the increase in the number of raters. On examining the categories constituting analytic rubric, it was found that there was variability in the levels of raters' agreement on the basis of objectivity. It was observed from the results that Kappa statistics and Krippendorff Alpha techniques yielded similar results. Moreover, Krippendorff alpha technique was found to be affected less by the number of raters. Log linear analysis technique, on the other hand, provided more comprehensive and extensive knowledge through showing the source of disagreement and interaction among the

*Bu makale, birinci yazar tarafından ikinci yazar danışmanlığında hazırlanan "Üst düzey zihinsel özelliklerin ölçülmesinde planlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması" başlıklı yüksek lisans tezinden üretilmiştir.

**Arş. Gör. Dr., Adnan Menderes Üniversitesi, Eğitim Fakültesi, Temel Eğitim Bölümü, Aydın-Türkiye, e-posta: ozgebikmaz@adu.edu.tr.

*** Prof. Dr., Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Eğitimde Ölçme ve Değerlendirme Bilim Dalı, Ankara-Türkiye, e-posta:nurid@hacettepe.edu.tr

Geliş Tarihi: 23.02.2017
DOI: 10.21031/epod.294847
Kabul Tarihi: 08.03.2017

variants. As a result, it is thought that analyzing the scores obtained by using the analytic rubric which is composed of sub-categories using log-linear analysis technique would be more appropriate when the purpose is to obtain more detailed measurement results whereas analyzing the scores obtained through holistic rubric by using the Krippendorff technique would be more appropriate when the purpose is to obtain more general results.

Keywords: Kappa statistic, log linear analysis technique, Krippendorff alpha

GİRİŞ

Ölçme ve değerlendirme eğitim sisteminin ayrılmaz ve önemli bir parçasıdır. Eğitim sisteminde ölçme ve değerlendirme uygulamalarından, eğitim sürecinin başında, süreç devam ederken ve sonunda yararlanılmaktadır. Hedeflenen istendik davranışların gerçekleşip gerçekleşmediğini belirleme, davranışta ne derecede bir değişim olduğunu saptama, süreçte kullanılan tekniklerin etkililik düzeylerini ve öğrenme sürecinde aksayan yönleri ortaya koyma amacıyla ölçme ve değerlendirme uygulamalarından yararlanılmaktadır. Ölçme ve değerlendirme sayesinde öğretim programının etkililiği değerlendirilebilir ve öğrencilerin öğrenme eksiklikleri ortaya konarak onların başarıları belirlenebilir (Atılgan, Kan ve Doğan, 2007).

Geleneksel ölçme yaklaşımları ile yapılan ölçmelerde, öğrencilerden, sınırlı bir zaman diliminde, kimseye danışmadan ya da belli kaynaklara başvurmadan verilen ölçme araçlarındaki soruları yanıtlaması beklenmektedir. Oysa bu tür bir ölçme sonucu gerçek yaşamda, öğrencilerin ölçme aracında verdiği yanıtlara uygun davrandığını kanıtlamadan uzaktır. Kutlu, Doğan ve Karakaya (2009) geleneksel değerlendirme yaklaşımlarının gerçek yaşam durumlarından uzak olmasının bir eleştiri noktası olduğunu, öğrencilerin çoktan seçmeli testlerle ya da boşluk doldurarak ne öğrendiklerinin değerlendirilmesinin çok zor olduğunu belirtmiştir. Geleneksel ölçme araçlarının belirtilen sınırlılıkları ile ölçülemeyen davranışlar performansa dayalı durum belirlemeyi gündeme getirmiştir.

Performansa dayalı durum belirleme Fitzpatrick ve Morrison (1971) tarafından "gerçek yaşam durumlarına benzer ortamlarda bireyin verdiği bir dizi yanıtın ölçülmesi" olarak tanımlanmaktadır. Bir dizi yanıt ile anlatılmak istenen, bireylerin ifade ettiği, yaptığı ya da ürettiğiyle ilgili davranışları içeren yanıtlardır. Bu bağlamda performanstan kasıt, bireyin ona sunulan öğrenme ortamında nasıl davrandığı ya da nasıl hareket ettiği ile ilgilenmektir. Eğitim alanında performans kavramıyla ilgili çok sayıda tanım yapılmıştır. Performansa dayalı durum belirleme süreci bağlamında performans, Kutlu ve diğerleri (2009) tarafından üst düzey zihinsel süreçleri gerektiren beceri ve yeteneklerle ilişkili olarak açıklanmıştır. Performans, beceri ve yetenekleri kapsayan karmaşık bir yapı olarak ele alınmaktadır. Performansa dayalı durum belirleme süreci bir dizi aşamayı gerektirmektedir. Bu aşamalar önceden belirlenmiş uygun bir görevin tanımlanması, yanıtlayıcıya verilmesi, yanıtlayıcı tarafından yerine getirilmesi ve sürecin gözlenerek puanlaması aşamalarını kapsamaktadır. Bu süreçte belirtilen puanlama aşamasında klasik ölçme araçlarından farklı olan ve performansın yapısına uygun ölçme araçları kullanılmaktadır.

Performansa dayalı durum belirleme sürecinde puanlama için kontrol listeleri ve puanlama ölçekleri kullanılmaktadır. Puanlama ölçeklerinden dereceli puanlama anahtarları bu süreçte sıklıkla kullanılan araçlarındandır (Airasian, 1994). Dereceli puanlama anahtarları, bir göreve ilişkin ölçüt listesini ve bu ölçütlere ilişkin niteliklerin derecesini içeren puanlama araçları olarak tanımlanmaktadır (Goodrich, 1997). Bu tür puanlama araçlarının, öğrencinin gelişmesi için dönüt sağlamak ya da öğrencinin belirli ölçütlere göre, ulaştığı düzeyi betimleyebilmek adına yararlı oldukları görülmektedir. Dereceli puanlama anahtarları alan yazında bütünsel (holistik) ve analitik olarak iki türde karşımıza çıkmaktadır (Mertler, 2001).

Analitik dereceli puanlama anahtarları, gözlemlere ait puanların, tanımlanmış kategorilerden (ölçütlerden) uygun düşen boyuta kaydedilmesini sağlayan ölçme araçlarıdır (Haladyna, 1997). Analitik puanlama anahtarı kullanımının ölçülecek performans çok boyutlu ve bileşenlerine ayrılabildiği durumlarda, performans görevine ilişkin öğretmene ve öğrenciye anlamlı geri bildirimler sağlaması istendiğinde, performansı belirlemek için yeterli süre olduğunda, üst eğitim kademelerinde kullanıldığı durumlarda yararlı oldukları belirtilmektedir (Mertler, 2001; Nitko, 2001). Bütünsel puanlama anahtarında ise analitik puanlama anahtarından farklı olarak öğrencinin gösterdiği

performans bütün olarak belirlenmekte ve öğrenciye tek bir puan verilmektedir (Kutlu ve diğerleri, 2009). Öğrencilerin performansı, parçalara ayrılmadan bütün olarak belirlenmek istendiği durumlarda kullanıldığı ifade edilmektedir (Moskal, 2000). Yani bu tür dereceli puanlama anahtarı öğrenci çalışmalarını bir bütün olarak ele alıp değerlendirmeyi amaçlamaktadır (Korkmaz, 2004). Bu nedenle bütünsel puanlama anahtarı performansın oluşumunda etkili olan süreçten çok ortaya çıkan ürüne, ürünün niteliğine ve sonucuna odaklanmaktadır (Atılgan, Kan ve Doğan, 2007).

Performansa Dayalı Durum Belirlemede Güvenirlik

Öğrencinin sergilediği performansın olduğu gibi kabul edilmesi olanaksızdır. Diğer ölçme araçlarında olduğu gibi performansın belirlenmesinde kullanılan araçların da geçerli ve güvenilir olması gerekmektedir. Güvenirlik, ölçme sonuçlarının tesadüfi hatalardan arınıklığı olarak tanımlanmaktadır (Baykul, 2000). Tesadüfi hata ölçme işlemine hangi kaynaktan, ne derece karıştığı belli olmayan hata türüdür. Ölçmeyi yapanın dikkatsizliği, ölçme ortamı, ölçme aracı ve diğer faktörler bu tür hataya neden olabilir (Atılgan, Kan ve Doğan, 2007). Cohen, Swerdlik ve Phillips (1996) bu tür hatayı ölçme aracının hazırlanması, uygulanması, puanlanması ve yorumlanması aşamalarında ölçmeye karışan istenmedik durumlar olarak açıklamıştır. Ölçmeye karışan bu istenmedik durumları saptamak, hatasız sonuçlar elde etmek için araştırmacılar hata kaynaklarının ölçme sonuçları üzerindeki etkilerini hesaplamaya yönelik güvenirlik belirleme tekniklerini önermişlerdir. Güvenirlik belirleme teknikleri, test tekrar test, eşdeğer formlar olarak iki uygulamaya dayalı yöntemler ile Cronbach Alfa, KR-20-21, iki yarı yöntemi gibi içtutarlılık katsayısı olarak ifade edilen ve tek uygulamaya dayalı yöntemler olarak sınıflanmaktadır (Crocker ve Algina, 1986). Performansa dayalı durum belirlemede puanlayıcılar önemli bir hata kaynağı olarak sayılmaktadır. İstenmeyen değişkenlik kaynağının puanlayıcılar olduğu performansa dayalı durum belirleme sürecinde güvenirlik kestirimi için puanlayıcılar arası güvenirlik belirleme teknikleri önerilmiştir (Cohen, Swerdlik ve Phillips, 1996). Diğer bir deyişle bu tür ölçme araçlarının güvenirliği puanlayıcı kanısına dayalı olarak elde edilmektedir.

Puanlayıcılar Arası Güvenirlik

Puanlayıcı güvenirliği puanlayıcı-içi ve puanlayıcılar-arası güvenirlik olarak iki türde incelenmektedir. Puanlayıcı-içi güvenirlik, aynı bireyin verdiği puanların birbiriyle tutarlılığı incelenerek hesaplanmaktadır. Çoğu araştırmada Cronbach Alfa katsayısı ile kestirilmektedir (Jonsson ve Svingby, 2007). Puanlayıcılar-arası güvenirlik ise birden fazla puanlayıcının verdiği puanlar arasındaki uyumun belirlenmesiyle hesaplanmaktadır. İki ya da daha fazla puanlayıcı (değerlendirici) arasındaki uyum veya tutarlılığın derecesi olarak tanımlanmaktır (Cohen ve diğerleri, 1996).

Puanlayıcılar arası güvenirlik aynı özelliği puanlayan birbirinden bağımsız iki ya da daha fazla puanlayıcı olduğunda herhangi bir durumda hesaplanabilmektedir (Viera ve Garret, 2005). Elde edilen güvenirlik değeri puanlayıcıların belli bir davranışın puanlanmasında ne derece fikir birliği içinde olduklarını yansıtmaktadır (Burry-Stock, Shaw, Laurie ve Chissom, 1996). Puanlayıcılar arası güvenirlik, puanlamanın bir puanlayıcıdan diğerine değişmemesi olarak tanımlanmaktadır (Kutlu ve diğerleri, 2009).

Puanlayıcılar Arası Güvenirlik Belirleme Teknikleri

Alan yazında puanlayıcılar arası güvenirlik belirlemede uyum yüzdesi, puanlayıcılar-arası korelasyon katsayısı, puanlar arasındaki farka dayalı ANOVA gibi çok sayıda teknik kullanıldığı görülmektedir (Jonsson ve Svingby, 2007). Güvenirlik belirlemede amaç tesadüfi hata kaynaklarını belirlemektir. Puanlayıcı güvenirliğinde tesadüfi hata kaynağı olarak puanlayıcılar ele alınmaktadır. Uyum yüzdesi ya da korelasyona dayalı hesaplamalarda puanlayıcıların tesadüfe dayalı uyumlulukları hesaplanmadığı için eleştirildiği çalışmalara rastlanmaktadır. Puanlayıcılar-arası güvenirlik kestirimi

için çeşitli avantajlara sahip çok sayıda teknik önerilmiş, tekniklerden Kappa istatistiği, Krippendorff Alfa katsayısı ve log-linear analiz teknikleri bu çalışma kapsamında incelenmiştir.

Kappa istatistiği (κ)

Puanlayıcılar arası güvenirlik belirlemede sıklıkla kullanılan Kappa istatistiği, Cohen (1960) tarafından önerilmiştir. Sınıflama düzeyinde puanlama yapan iki puanlayıcı arasındaki uyumun derecesini belirlemek için geliştirilmiştir (Cohen, 1960). İki puanlayıcı ile sınırlı kalan κ istatistiği, Fleiss (1971) tarafından ikiden fazla puanlayıcı arasındaki uyumu belirlemede kullanılabilmesi için genelleştirilmiştir (Fleiss, 1971). Kappa istatistiği bazı temel varsayımlara dayanmaktadır (Crawforth, 2001). Bu varsayımlar Brennen ve Prediger (1981) tarafından puanlama sürecinde kategorilenen nesne ya da bireylerin bağımsız olduğu, puanlayıcıların puanlamalarının birbirinden bağımsız olduğu, puanlamada kullanılan kategorilerin birbirinden bağımsız olduğu şeklinde ifade edilmiştir. Kappa istatistiğinin bir avantajı kolay hesaplanması ve pratik yorumlanmasıdır. Diğer ve en önemli avantajı ise şansla beklenen uyumu düzeltmeyi temel almasıdır. Şansla meydana gelen uyum puanlardaki tamamen tesadüfe dayalı oluşan benzerliktir. κ, puanlayıcılar arası gözlenen uyumun içinden şansa/tesadüfe dayalı uyumun çıkarılmasına dayalı olarak Eşitlik 1'de verilen formülle hesaplanmaktadır. $\bar{P}$ gözlenen uyumluluk oranı, $\bar{P}_e$ tesadüfî/şansla uyumluluk oranı olmak üzere kappa istatistiği (κ) formülüyle hesaplanmaktadır (Sim ve Wright, 2005).

$$\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e} \qquad \textit{(Eşitlik 1)}$$

Kappa istatistiği -1 ile +1 arasında değer almaktadır (Fleiss, 1971). κ'nın pozitif değerleri puanlayıcılar arasındaki uyumun şansla beklenen uyumdan daha fazla olduğunu, κ'nın negatif değerleri puanlayıcılar arasındaki uyumun şansla beklenenden daha az olduğunu göstermektedir (Von Eye ve Mun, 2005). Bu anlamda negatif değerler şansla beklenenin altındaki uyum düzeyini gösterdiği için dikkate alınmamaktadır (Goodwin, 2001). κ istatistiğinin yorumlanmasında Tablo 1'de Landis ve Koch (1977) tarafından önerilen uyum düzeyleri kullanılmaktadır.

Tablo 1. Kappa İstatistiğinin Yorumlanmasına İlişkin Değer Aralıkları

κ	Uyumun Gücü
< 0,00	Zayıf
0,00 – 0,20	Önemsiz
0,21 – 0,40	Düşük
0,41 – 0,60	Orta
0,61 – 0,80	Önemli
0,81 – 1,00	Çok Yüksek

Krippendorff Alfa katsayısı (α)

Krippendorff (1995) tarafından Krippendorff Alfa istatistiği adlı bir uyum ölçüsü önerilmiştir. Bu katsayı ilk olarak içerik analizinde kodlayıcılar arasındaki uyumun ölçüsünü belirlemeye yönelik olarak geliştirilmiştir. Bir uyum istatistiği olarak puanlayıcılar arasındaki uyumu belirlemede de kullanılmaktadır (Krippendorff, 1995, 2004, 2007). Krippendorff Alfa (α) istatistiği çok çeşitli veri tiplerine uygulanabilmektedir. Her değişken için herhangi bir sayıdaki değere uygulanabilir. İki veya daha fazla puanlayıcı içeren verilere uygulanabilir. Herhangi bir ölçek türü (sınıflama, sıralı, aralık, oran) ile ölçülmüş verilere uygulanabilir ve farklı büyüklükteki (küçük veya büyük) örneklemlerde kullanılabilir. Ayrıca puanlamada eksik veri olduğu durumlarda da uygulanabilir (Krippendorff, 1995). Alfa istatistiği için öncelikle gözlenen uyumsuzluk (D_0) ve beklenen uyumsuzluk (D_e) hesaplanmaktadır. Gözlenen uyumsuzluğun beklenen uyumsuzluğa bölünmesiyle elde edilen değer'in 1'den çıkartılması sonucunda α elde edilmektedir. Krippendorff Alfa katsayısının formülü Eşitlik 2'de verildiği şekildedir:

$$\alpha = 1 - \frac{D_0}{D_e} \qquad \textbf{(Eşitlik 2)}$$

Krippendorff alfa istatistiğinin yorumlanmasında α=1 olması puanlayıcılar arasındaki uyumun mükemmel olduğunu, α= 0 ise tam uyumsuzluğu simgelemektedir. Şansa bağlı olarak puanlayıcılar uyumlu oldukları zaman $D_o = D_e$ (Yani α=0) olur. Bu durum uyumun olmadığına işaret etmektedir. α'nın negatif çıkması şansla beklenin altında bir uyum olduğunu göstermektedir. Yüksek düzeyde güvenirliğin elde edilmesi amaçlandığından yorumlamada negatif değerler dikkate alınmamaktadır. α istatistiğinin yorumlanmasında Tablo 2'de verilen Krippendorff (1995) tarafından önerilen uyum düzeyleri kullanılmaktadır.

Tablo 2. Krippendorff Alfa Katsayısının Yorumlanmasına İlişkin Değer Aralıkları

α	Uyumun gücü
< 0,67	Zayıf
0,67 - 0,80	Orta
0,80 ≤	Yüksek

Log-linear Analiz Tekniği

Çok yönlü çapraz tablolardaki kategorik veriler arasındaki ilişkiyi analiz etmek için ki-kare istatistiğinin uygulanabildiği ve yetersiz kaldığı durumlarda çok yönlü tabloların analizini modeller aracılığıyla yapabilen bir teknik olarak önerilmiştir (Agresti, 1996). Puanlayıcılar arası güvenirlik belirleme tekniği olarak kullanımı Tanner ve Young (1985)'in bu tekniği puanlayıcılar arasındaki uyumu modelleme denemeleriyle alan yazına girmiştir. Diğer tekniklerden farklı olarak sıralı ya da gruplanarak kategorik hale dönüştürülen eşit aralık ve oran ölçeğindeki verilerin iki yönlü, çok yönlü ve iç içe çapraz tablolarında birlikte değişimleri ve değişkenlerin alt kategorileri ile arasındaki etkileşimlerini analiz etmede kullanılmaktadır. Bu teknikte gözlenen frekansların modelini oluşturmak önemlidir. Seçilen modele göre her bir kategori için frekanslar hesaplanır ve bunlar gözlenen frekanslarla karşılaştırılır. Eğer uyum yeterli değilse model red edilir. Diğer bir deyişle log-linear analizde k tane değişken analize alınıyorsa "k veya daha fazla dereceli etkiler 0'a eşittir" hipotezine yanıt aranır. Elde edilen değer istatistiksel açıdan önemliyse H_0 red edilir.

Log-linear analiz sonuçlarının yorumlanmasında modelin uygun olup olmadığına Eşittlik 3 ve 4'te verilen Pearson ki-kare (χ^2) ve olasılık oranı L^2 ile karar verilmektedir.

$$\chi^2 = \sum \frac{(O_i - E_i)}{E_i} \qquad \textbf{(Eşitlik 3)}$$

ve

$$L^2 = 2 \sum O_i \ln\left(\frac{O_i}{E_i}\right) \qquad \textbf{(Eşitlik 4)}$$

Yukarıdaki formüllerde, O_i, gözlenen hücre frekansı, E_i, beklenen hücre frekansıdır. Eğer sonuç istatistiksel açıdan önemliyse, model uyumu zayıftır, H_0 hipotezi, dolayısıyla da model red edilir (Agresti ve Yang, 1987).

Yukarıda açıklanan tekniklerin avantajları ve sınırlılıkları mevcuttur. Bu çalışmada ilgili teknikler üzerinde dereceli puanlama anahtarı türü ve puanlayıcı sayısı değişiminin etkisinin incelenmesi için iki alt problem oluşturulmuştur.

1. İki, beş ve on puanlayıcı ile analitik dereceli puanlama anahtarı aracılığıyla puanlama yapıldığında, Kappa istatistiği, Krippendorff Alfa katsayısı ve Log linear analiz tekniği ile elde edilen güvenirlik sonuçları nasıldır?

2. İki, beş ve on puanlayıcı ile bütünsel puanlama anahtarı aracılığıyla puanlama yapıldığında, Kappa istatistiği, Krippendorff Alfa katsayısı ve Kendall uyum istatistiği ile elde edilen güvenirlik sonuçları nasıldır?

Araştırmanın Amacı

Araştırmanın amacı, aynı amaca yönelik hazırlanan ve aynı bireylere uygulanan performansın analitik ve bütünsel puanlama anahtarı kullanılarak puanlanmasıyla Kappa, Krippendorf ve Log-Linear analiz teknikleriyle hesaplanan güvenirlik değerlerini incelemek ve karşılaştırmaktır. Aynı zamanda bu araştırmada, puanlayıcı sayısı değişiminin, tekniklere ve ölçme aracına bağlı olarak puanlayıcı güvenirliğinde bir değişime neden olup olmadığının belirlenmesi amaçlanmaktadır.

YÖNTEM

Araştırmanın Türü

Bu araştırma, Kappa istatistiği, Krippendorff Alfa katsayısı ve log-linear analiz tekniğinin uygulanmasına, bu tekniklerin benzerlik ve farklılıklarının belirlenmesine, sınırlılıklarının incelenmesine, tekniklerden hangisinin daha fazla bilgi sağladığının saptanmasına dayanmaktadır. Bu yönüyle durum saptamaya yönelik olduğu için betimsel bir araştırma niteliği taşımaktadır.

Çalışma Grubu

Araştırmanın çalışma grubu, performans görevini yerine getiren 50 öğrenciden ve bu görevleri biri analitik diğeri bütünsel iki ayrı dereceli puanlama anahtarı kullanarak puanlayan 10 öğretmenden oluşmaktadır. Çalışma grubunda yer alan öğrenciler, araştırmacılardan ilkinin sınıf öğretmeni olarak görev yaptığı İstanbul ili Beyoğlu ilçesine bağlı bir devlet okulunun beşinci sınıfında öğrenim gören 50 öğrenciden oluşmaktadır. Öğrencilerin belirlenmesinde, öğrencilere kolay ulaşabilme, uygulama döneminde öğrencileri izleyebilme gibi kolaylıklar etkili olmuştur. Çalışma grubunda yer alan öğretmenler, Milli Eğitim Bakanlığı yapısındaki İstanbul ilinde sınıf öğretmeni olarak görev yapan öğretmenlerden oluşmaktadır. Öğretmenlerin seçiminde gönüllük esası benimsenmiştir.

Veri Toplama Araçları

Performans görevi

İlköğretim fen ve teknoloji dersi öğretim programında yer alan "Vücudumuzun bilmecesini çözelim" ünitesiyle ilişkili Kutlu, Doğan ve Karakaya (2009) tarafından hazırlanan performans görevidir.

Analitik dereceli puanlama anahtarı

Kutlu ve diğerleri (2009) tarafından geliştirilen analitik puanlama anahtarından yararlanılarak oluşturulmuştur. Analitik puanlama anahtarı altı alt kategoriden oluşmaktadır. Bu kategoriler: içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma ve zaman kullanımıdır. Beş ölçme ve değerlendirme uzmanı kanısı, çalışma koşulları düşünülerek zaman kullanımı alt kategorisi analize dahil edilmeden son şekliyle analitik puanlama anahtarı beş alt boyutlu olarak oluşturulmuştur. Her bir performans 1-3 arasında puanlanmaktadır. Alınabilecek en yüksek puan 15; en düşük puan 5'tir.

Bütünsel dereceli puanlama anahtarı

Kutlu ve diğerleri (2009) tarafından geliştirilen analitik puanlama anahtarından yararlanılarak oluşturulmuş, beş ölçme ve değerlendirme uzmanı kanılarına dayalı olarak son şekli verilmiştir. Bu araçtan alınabilecek en yüksek puan 4 iken en düşük puan 1'dir.

İşlem

Çalışma kapsamında işlemler, performans görevinin uygulanması ve bu görevin puanlanması olarak iki aşamada gerçekleştirilmiştir.

Performans görevinin uygulanması

Öğrencilere performans görevi ve performanslarının değerlendirileceği ölçütler performans göreviyle birlikte yazılı olarak verilmiş, varsa anlamadıkları noktalar öğrencilere açıklanmıştır. Performans görevini tamamlamaları için öğrencilere 10 gün süre verilmiştir. 10 gün boyunca araştırmacının belirlediği zamanda öğrencilerle görüşülmüş, varsa soruları yanıtlanmıştır. Öğrencilere verilen süre sonlandığında performans görevlerine ilişkin raporlar toplanmış, görevlere araştırmacı tarafından puan verilmiştir.

Performans görevinin puanlanması

Puanlamayı yapan öğretmenlere performans görevi, dereceli puanlama anahtarları ve puanlama süreciyle ilgili genel bilgi verilmiştir. Puanlamada hatırlama etkisini ortadan kaldırmak adına öğrencilere 1'den 50'ye kadar numaralar verilerek kod numaraları oluşturulmuştur. Her öğretmen ayrı bir zaman diliminde puanlama yapmıştır. Puanlamada yine hatırlama etkisini ortadan kaldırmak adına her bir öğretmenin bütünsel dereceli puanlama anahtarı kullandığı ikinci puanlama ile ilk puanlama arasında 2 hafta süre bırakılmıştır. Puanlayıcıların aynı oturumda tüm öğrencilere puan verme işlemini tamamlaması sağlanmıştır.

Verilerin Analizi

Verilerin çözümlenmesinde analitik puanlama anahtarı için puanlayıcılar arası güvenirlik belirleme tekniklerinden Kappa istatistiği, Krippendorff alfa katsayısı, log linear analiz tekniği kullanılmıştır. Log-linear analiz sadece kategorik verilere uygulanabildiği için bütünsel dereceli puanlama anahtarı ile elde edilen sonuçların analizinde kullanılamamıştır. Bütünsel dereceli puanlama anahtarı ile verilerin analizi için log-linear analiz yerine Kendall'ın uyum istatistiği kullanılmıştır. Bu haliyle bütünsel puanlama anahtarından elde edilen veriler için Kappa istatistiği, Krippendorff alfa katsayısı ve Kendall'ın uyum katsayısından yararlanılmıştır. Kappa istatistiğinin hesaplanmasında "SPSS syntax (mkappasc.sps)" dosyasından, Krippendorff için "SPSS syntax (kalpha.sps)" dosyasından, log linear analiz tekniği ve Kendall katsayıları için ise doğrudan SPSS paket programından yararlanılmıştır.

BULGULAR

Araştırmadan elde edilen bulgulara dayalı olarak ulaşılan sonuçlar araştırma soruları doğrultusunda değerlendirilmiştir. Analitik dereceli puanlama anahtarı kullanılarak yapılan puanlamada Kappa istatistiğine ilişkin bulgular Tablo 3'te özetlenmiştir.

Tablo 3. Analitik Puanlama Anahtarı ile Yapılan Puanlamaların Kappa İstatistiğiyle Hesaplanan Güvenirlik Değerleri

Puanlama Yapan Puanlayıcı Sayısı	Puanlama Anahtarının Kategorileri	Kappa İstatistiği Değeri (κ)
2	İçerik	0,40*
	Araştırma Süreci	0,59*
	Materyal Kullanımı	0,65*
	Grafik Oluşturma	0,81*
	Tablo Oluşturma	0,92*
5	İçerik	0,34*
	Araştırma Süreci	0,57*
	Materyal Kullanımı	0,64*
	Grafik Oluşturma	0,78*
	Tablo Oluşturma	0,84*
10	İçerik	0,32*
	Araştırma Süreci	0,52*
	Materyal Kullanımı	0,64*
	Grafik Oluşturma	0,75*
	Tablo Oluşturma	0,81*

**p<0,001*

Tablo 3'e göre iki, beş ve on puanlayıcının verdikleri puanlar arasındaki uyumu elde etmek amacıyla analitik puanlama anahtarının kategori değişkeninin her biri için hesaplanan Kappa değerleri istatistiksel açıdan anlamlı bulunmuştur ($p<0,001$). Burada κ istatistiklerinin pozitif olması, puanlayıcılar arasında tesadüfen çıkabilecek olası uyumdan daha yüksek düzeyde uyum olduğuna işaret etmektedir (Landis ve Koch, 1977). İki puanlayıcının olduğu koşul için Kappa değerleri 0,40 ile 0,92 arasındadır. Bu koşulda en düşük uyum, içerik kategorisinde elde edilmişken ($\kappa = 0,40$); en yüksek uyum, tablo oluşturma kategorisi için kestirilmiştir ($\kappa = 0,92$). Diğer üç kategori değerlendirildiğinde, araştırma süreci kategorisinde puanlayıcılar arasında orta düzeyde, materyal kullanımında önemli düzeyde, grafik oluşturmada ise çok yüksek düzeyde uyum olduğu görülmektedir.

Tablo 3'teki beş puanlayıcının olduğu koşul için κ istatistiğine ait değerlerin, 0,34 ile 0,84 arasında; on puanlayıcı olduğu koşulda ise 0,32 ile 0,81 arasında değişkenlik gösterdiği görülmektedir. İki, beş ve on puanlayıcının olduğu bulgular karşılaştırıldığında, Kappa istatistiği değerlerinin anlamlılık düzeyi ve değerlerinin kategori bazında κ değerlerinin büyükten küçüğe doğru sıralanışı değişmemiştir. Ancak Kappa değerleri puanlayıcı sayısı arttıkça görece azalmıştır. Tablo 3'te verildiği gibi analitik puanlama anahtarının kategorilerinin κ değerleri sırayla içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma olarak artan değerler almıştır. "içerik" kategorisinden "tablo oluşturma" kategorisine doğru gidildikçe elde edilen Kappa istatistiklerinin değeri artmıştır. "Tablo oluşturma" kategorisinde öğrenciden oluşturacağı tabloda satır, sütun isimlerini vermesi ve gözeneklerde bilgi vermesi beklenmektedir. "İçerik" kategorisinde ise öğrenciden konuyla ilgili tanıtıcı, açıklayıcı ve kaynaklara dayalı bilgi vermesi istenmektedir. Bu kategoride öğrenciden beklenen performansın çerçevesi puanlayıcının konuyla ilgili bilgi düzeyine, konuyu algılayışına ve yorumlayışına göre değişebilir. Nitekim bazı puanlayıcılar öğrencilerden konuyla ilgili performansın ayrıntılı bir sunumunu, bazıları ise kısa ve net sunumunu beklemektedir. Puanlayıcıların nitelikli buldukları sunumlar kişiden kişiye değişmektedir. Yani kategoride yer alan ölçüt bir yönüyle puanlayıcıdan puanlayıcıya farklılık gösterebilmektedir. Hesaplanan Kappa istatistiği değerlerinin Tablo 3'teki sırasıyla "içerik" kategorisinden, "tablo oluşturma" kategorisine gidildikçe artması, "tablo oluşturma" kategorisine doğru kategorilerin daha objektif olmasına dayandırılabilir.

Araştırmanın ilk alt problemi kapsamında analitik dereceli puanlama anahtarı ile puanlama yapıldığında Krippendorff alfa istatistiği ile iki, beş ve on puanlayıcı olduğunda elde edilen puanlamaların güvenirliğine ait bulgular Tablo 4'te özetlenmiştir.

Tablo 4. Analitik Puanlama Anahtarı ile Yapılan Puanlamaların Krippendorff Alfa Katsayısıyla Hesaplanan Güvenirlik Değerleri

Puanlama Yapan Puanlayıcı Sayısı	Puanlama Anahtarının Kategorileri	Alfa Katsayısı Değeri (α)
2	İçerik	0,59
	Araştırma Süreci	0,74
	Materyal Kullanımı	0,79
	Grafik Oluşturma	0,87
	Tablo Oluşturma	0,93
5	İçerik	0,52
	Araştırma Süreci	0,75
	Materyal Kullanımı	0,78
	Grafik Oluşturma	0,86
	Tablo Oluşturma	0,89
10	İçerik	0,51
	Araştırma Süreci	0,71
	Materyal Kullanımı	0,78
	Grafik Oluşturma	0,86
	Tablo Oluşturma	0,87

Analitik puanlama anahtarı kullanılarak yapılan puanlamada, puanlama anahtarının kategorileri için hesaplanan Krippendorff Alfa katsayısına ilişkin bulgular Tablo 4'te verilmiştir. Tablo 4 incelendiğinde, iki puanlayıcının puanları arasındaki uyumu elde etmek için kestirilen Krippendorff alfa katsayısı değerleri 0,59 ile 0,93 arasındadır. En düşük uyum "içerik" kategorisinde çıkmışken (α= 0,59); en yüksek uyum "tablo oluşturma" kategorisi için elde edilmiştir (α= 0,93). "Araştırma süreci" ve "materyal kullanımı" kategorileri için orta düzeyde uyum; "grafik oluşturma" ve "tablo oluşturma" kategorileri için yüksek düzeyde uyum olduğu tespit edilirken, içerik kategorisinde zayıf uyum olduğu görülmektedir. Buna göre "içerik" dışındaki kategorilerden elde edilen puanların güvenilir olduğu söylenebilir.

Beş puanlayıcının puanları arasındaki uyumu için hesaplanan Krippendorff alfa değerleri 0,52 ile 0,89 arasındadır. En düşük uyum "içerik" kategorisinde (α = 0,52); en yüksek uyum "tablo oluşturma" kategorisi için elde edilmiştir (α = 0,89). On puanlayıcının olduğu koşullarda alfa değerleri 0,51 ile 0,87 arasında değişmiştir. İki, beş ve on puanlayıcının olduğu bulgular karşılaştırıldığında, Krippendorff alfa katsayısı değerlerinin kategori bazında büyükten küçüğe doğru sıralanışı (tablo oluşturma, grafik oluşturma, materyal kullanımı, araştırma süreci, içerik) şeklinde değişmemiştir. Ancak alfa değerleri puanlayıcı sayısı ikiden beşe, beşten ona çıkarıldığında görece azalmıştır.

Analitik dereceli puanlama anahtarı kullanılarak yapılan puanlamada log-linear analiz bulgularına ait değerler Tablo 5'te verilmiştir.

Tablo 5. Analitik Puanlama Anahtarı ile Yapılan Puanlamaların Log-linear Analiz Tekniğiyle Hesaplanan Güvenirlik Değerleri

Puanlama Yapan Puanlayıcı Sayısı	Puanlama Anahtarının Kategorileri	*LL* χ^2	Etki Düzeyi
2	İçerik	34,072*	1. Düzey
	Araştırma Süreci Materyal Kullanımı Grafik Oluşturma	56,233*	2. Düzey
	Tablo Oluşturma	7,759	3. Düzey
5	İçerik	55,931*	1. Düzey
	Araştırma Süreci Materyal Kullanımı Grafik Oluşturma	144,333*	2. Düzey
	Tablo Oluşturma	36,636	3. Düzey
10	İçerik	34,616*	1. Düzey
	Araştırma Süreci Materyal Kullanımı Grafik Oluşturma	314,451*	2. Düzey
	Tablo Oluşturma	79,774	3. Düzey

*$p<0,05$

Log-linear analiz tekniğinde puanlayıcılar, kategoriler, alt kategoriler şeklinde üç değişken tanımlanmıştır. Öncelikle ilgili değişkenlerin tek başına etkilerinin yanında ikili ve üçlü etkilerinin birlikte bulunduğu logaritmik modelle ifade edilip edilemeyeceği incelenmiştir.

Kategoriler değişkeni analitik puanlama anahtarındaki "içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma" olarak beş tanedir. Kategoriler değişkenin her birine 1 ile 3 arasında değişen puan verilerek alt kategoriler değişkeni tanımlanmıştır. Diğer bir ifadeyle alt kategoriler değişkeni kategoriler değişkenine verilen puanlara dayanarak oluşturulmuştur. Her kategori değişkeninin altında üç tane altkategori değişkeni mevcuttur. İki, beş ve on puanlayıcı olduğu durumlarda bu değişkenlerin tek başına, ikili ve üçlü etkilerinin birlikte bulunduğu koşulların logaritmik modelle ifade edilip edilemeyeceği sınanmış, tekli ve ikili etkilerin istatistiksel açıdan anlamlı ($p<0,05$); üçlü etkilerin ise anlamsız olduğu ($p>0,05$) görülmüştür. Anlamlı olan tekli ve ikili etkiler için gerçekleştirilen log-linear analiz bulguları Tablo 6'da verilmiştir.

Tablo 6. Log-linear Analiz Tekniğinden Elde Edilen Bulgulara İlişkin Değerler

Puanlama Yapan Puanlayıcı Sayısı	Değişkenlerin Tekli ve İkili Etkileri	Serbestlik Derecesi	*Kısmi* χ^2
2	Puanlayıcı*kategori	4	0,026
	Puanlayıcı*altkategori	2	0,486
	Altkategori*kategori	8	55,791*
	Puanlayıcı	1	0,000
	Kategori	4	0,000
	Altkategori	2	34,072*
5	Puanlayıcı*kategori	16	0,680
	Puanlayıcı*altkategori	8	11,693
	Altkategori*kategori	8	133,320*
	Puanlayıcı	4	0,000
	Kategori	4	0,000
	Altkategori	2	55,931*
10	Puanlayıcı*kategori	36	2,120
	Puanlayıcı*altkategori	18	50,054
	Altkategori*kategori	8	266,517*
	Puanlayıcı	9	0,000
	Kategori	4	0,000
	Altkategori	2	34,616*

*$p<0,05$

Tablo 6'da yer alan bulgular incelendiğinde, iki puanlayıcının olduğu koşulda, değişkenlerin tek başlarına etkilerinin anlamsız olduğu yönünde kurulan hipotezlerden puanlayıcı ve kategori değişkenleri için olanlar kabul edilmiştir. Diğer bir ifade ile bu değişkenlerde puanlayıcılar arasındaki farklılık manidar değildir ($p>0,05$). İçerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma şeklinde beş kategoriden oluşan kategoriler bazında puanlamalarda manidar farklılık yoktur. "Altkategori" değişkenin etkisinin anlamsız olduğu yönündeki hipotez testi ise red edilmiştir. Yani alt kategoriler bazındaki puanlamalar arasında anlamlı fark vardır ($p<0,05$). İki puanlayıcının olduğu durumda tekli etkilere göre bulgular puanlamalardaki farklılığın "altkategori" değişkeninden kaynaklanabileceğini göstermektedir ($\chi^2 = 34,07$; sd= 2; $p<0,05$).

Puanlayıcı sayısının iki olduğu koşulda, puanlamada değişkenlerin ikili etkileşimlerinin farklılıklarının anlamsız olduğu yönündeki hipotezlerden puanlayıcı*altkategori ve puanlayıcı*kategori etkileşimleri için kurulanlar kabul edilmiştir. Buna göre bu etkileşimlerin anlamlı olmadığı görülmektedir ($p>0,05$). Ancak altkategori*kategori etkileşiminin anlamsız olduğu yönünde kurulan hipotez red edilmiştir, yani bu etkileşimin puanlamada oluşturduğu farklılık anlamlı bulunmuştur ($p<0,05$). "Altkategori" değişkeninin hem tek başına hem de kategori değişkeniyle etkileşimi sonucunda farklılık oluşturması puanlardaki asıl değişkenlik kaynağının "altkategori" değişkeni olduğu şeklinde yorumlanabilir ($\chi^2 = 55,79$; sd=8; $p<0,05$).

Tablo 6'da beş puanlayıcı için log linear analiz bulguları incelendiğinde, puanlayıcı ve kategori değişkenlerinin tek başına etkilerinin anlamlı olmadığı; "altkategori" etkisinin ise anlamlı olduğu tespit edilmiştir ($\chi^2 = 55,931$; sd= 2; $p<0,05$). Değişkenlerin ikili etkileşimleri incelendiğinde ise puanlayıcı*altkategori ve puanlayıcı*kategori etkilerinin anlamsız; kategori*altkategori etkisinin ise anlamlı olduğu gözlenmiştir ($\chi^2 = 133,32$; sd=8; $p<0,05$). "Altkategori" değişkeninin hem tek başına hem de kategori değişkeniyle etkileşimi sonucunda farklılık oluşturması puanlardaki asıl değişkenlik kaynağının "altkategori" değişkeni olduğu şeklinde yorumlanabilir.

Tablo 6'da on puanlayıcı için log linear analiz bulguları incelendiğinde, puanlayıcı ve kategori değişkenlerinin tek başına etkilerinin anlamlı olmadığı; "altkategori" etkisinin ise anlamlı olduğu tespit edilmiştir ($\chi^2 = 34,616$; sd= 2; $p<0,05$). Değişkenlerin ikili etkileşimleri incelendiğinde ise puanlayıcı*altkategori ve puanlayıcı*kategori etkilerinin anlamsız; altkategori*kategori etkisinin ise anlamlı olduğu gözlenmiştir ($\chi^2 = 266,517$; sd=8; $p<0,05$). "Altkategori" değişkeninin hem tek başına hem de kategori değişkeniyle etkileşimi sonucunda farklılık oluşturması, puanlardaki asıl değişkenlik kaynağının "altkategori" değişkeni olduğu şeklinde yorumlanabilir.

Bütünsel dereceli puanlama anahtarı kullanılarak yapılan puanlamada Kappa istatistiğine ilişkin bulgular Tablo 7'de özetlenmiştir.

Tablo 7. Bütünsel Puanlama Anahtarı ile Yapılan Puanlamaların Kappa İstatistiğiyle Hesaplanan Güvenirlik Değerleri

Puanlama Yapan Puanlayıcı Sayısı	Kappa İstatistiği Değeri (κ)
2	0,38*
5	0,24*
10	0,27*

*$p<0,001$

Tablo 7'ye göre iki puanlayıcının bütünsel dereceli puanlama anahtarı ile yaptıkları puanlamaların uyumu için kullanılan Kappa istatistiğinin değeri 0,38; beş puanlayıcı için 0,24; on puanlayıcı için ise 0,27 olarak hesaplanmıştır. Buna göre bütünsel puanlama anahtarı ile yapılan puanlamalarda Kappa değeri ile puanlayıcıların birbiriyle düşük düzeyde uyum gösterdikleri görülmektedir.

Bütünsel dereceli puanlama anahtarı kullanılarak yapılan puanlamalar arasındaki uyum için hesaplanan Krippendorff alfa istatistiğine ilişkin bulgular Tablo 8'de özetlenmiştir.

Tablo 8. Bütünsel Puanlama Anahtarı ile Yapılan Puanlamaların Krippendorff İstatistiğiyle Hesaplanan Güvenirlik Değerleri

Puanlama Yapan Puanlayıcı Sayısı	Alfa Katsayısı Değeri (α)
2	0,67
5	0,60
10	0,58

Tablo 8'e göre iki puanlayıcının bütünsel puanlama anahtarı ile verdiği puanlar arasındaki Krippendorf alfa istatistiği ile hesaplanan uyum değerleri 0,67; beş puanlayıcı ile hesaplanan uyum 0,60; on puanlayıcı ile hesaplanan uyum değerleri ise 0,58 olarak belirlenmiştir. Bu değerler, puanlayıcı arasındaki uyumun zayıf düzeyde olduğunu göstermektedir.

Tablo 9. Bütünsel Puanlama Anahtarı ile Yapılan Puanlamaların Kendall Uyum Katsayısıyla Hesaplanan Güvenirlik Değerleri

Puanlama Yapan Puanlayıcı Sayısı	Kendall Katsayısı Değeri (w)
2	0,61*
5	0,31*
10	0,18*

*$p<0,001$

Tablo 9'a göre iki puanlayıcının puanları arasındaki uyumu elde etmek için hesaplanan Kendall uyum istatistiği 0,61 olarak hesaplanmış ve bu değer istatistiksel açıdan anlamlı bulunmuştur ($p<0,001$). Sıra farklarını dikkate alarak uyumun hesaplandığı bir teknik olan Kendall'ın uyum istatistiği iki puanlayıcı için orta düzeyde çıkmıştır. Bu bulgu puanlayıcıların bireyleri sıralamada farklılık gösterdiği şeklinde yorumlanabilir. Beş puanlayıcının puanları arasındaki uyumu elde etmek için kullanılan Kendall'ın uyum istatistiği 0,31 olarak hesaplanmış ve bu değer anlamlı çıkmıştır ($p<0,001$). Bu değerin, Von Eye ve Mun (2005)'un ölçütlerine göre düşük düzeyde olduğu söylenebilir. On puanlayıcının puanları arasındaki uyumu elde etmek için hesaplanan Kendall'ın uyum istatistiği istatistiksel açıdan anlamlı bulunmuştur ($p<0,001$). Hesaplanan değer 0,18 olup, bu değer uyumun zayıf düzeyde olduğunu göstermektedir. Sıra farklarını dikkate alarak uyumun hesaplandığı teknik olan Kendall istatistiğinin çok düşük olması puanlayıcıların bireyleri sıralamada farklılık gösterdiği şeklinde yorumlanabilir.

SONUÇLAR ve TARTIŞMA

Aynı amaca yönelik analitik puanlama anahtarı ile yapılan puanlama, bütünsel puanlama anahtarı kullanılarak yapılan puanlamaya göre göreceli olarak daha objektif sonuçlar vermiştir. Puanlayıcılar arasında daha standart ve daha nesnel sonuçlar veren analitik dereceli puanlama anahtarının daha tutarlı puanlama sağladığı, dolayısıyla daha güvenilir olduğu sonucuna varılmıştır. Bu sonuç, Kutlu ve diğerleri (2009)'nin bütünsel dereceli puanlama anahtarından elde edilen sonuçların analitik puanlama anahtarından elde edilen sonuçlara göre güvenirlik düzeyinin düşük olduğu yönündeki açıklamalarıyla örtüşmektedir. Aynı zamanda bu çalışmanın sonuçları, Jonsson ve Svingby (2007)'ın çalışmasının analitik puanlama anahtarının güvenirliği artırdığı yönündeki sonuçlarını desteklemektedir.

Araştırma kapsamında, alt kategorilere sahip analitik puanlama anahtarı kullanılarak elde edilen sonuçlar incelenmiştir. Puanlama anahtarını oluşturan kategorilerin objektifliğinin, içerik, araştırma süreci, materyal kullanımı, grafik oluşturma, tablo oluşturma sırasıyla arttığı görülmüştür. Bu

bulgudan yola çıkarak, analitik puanlama anahtarı kullanımının puanlayıcılar arasındaki farklılıkları tamamen ortadan kaldırmada yeterli olmadığı; fakat puanlamalar arasında tutarlılığı arttırarak objektiflik düzeyini arttırdığı söylenebilir.

Araştırmada, hem analitik puanlama hem de bütünsel puanlama anahtarından elde edilen puanların üç teknikle yapılan analizlerinde en yüksek güvenirlik değerleri iki puanlayıcı olduğu durumda elde edilmiş, puanlayıcı sayısı arttıkça güvenirliğin giderek düştüğü sonucuna varılmıştır. Bu sonuç Abedi, Baker ve Herl (1995)'in çalışmalarının performansın ölçülmesinde puanlayıcı sayısı artışının puanlardaki değişkenlik düzeyini arttırarak güvenirliği düşürdüğü bulgusuyla ve Nying (2004)'in güvenirliğin puanlayıcı sayısı artışından etkilendiği bulgusuyla örtüşmektedir.

Araştırma kapsamında, Kappa istatistiği tekniğiyle puanlayıcı sayısına dayalı olarak yapılan üç analizde en yüksek uyum iki puanlayıcı olduğu koşulda elde edilmiştir. Kappa istatistiği tekniğinde puanlayıcı sayısının artışı, Kappa değerini düşürmüş; ancak istatistiğin anlamlılık düzeyi aynı kalmıştır. Bu durum Kappa istatistiğinin puanlayıcı sayısından etkilendiğinin göstergesi olabilir. Araştırmanın bu sonucu Nying (2004)'in Kappa istatistiğinin puanlayıcı sayısından olumsuz etkilendiği bulgusuyla paraleldir.

Araştırma sonucunda Krippendorff alfa tekniğiyle yapılan üç analizden en yüksek değerler, iki puanlayıcı arasındaki uyuma ilişkin çıkmıştır. Puanlayıcı sayısının artışı alfa değerini değiştirmiş; ancak bu değişim Kappa istatistiğindeki kadar değişkenlik göstermemiş ve daha kararlı yapı sergilemiştir. Analizde puanlayıcı sayısı ikiden beşe çıkarıldığında alfa değerinde düşüş yaşanmıştır; ancak puanlayıcı sayısı ona çıkarıldığında alfa değerinde önemli düzeyde düşüş olmamıştır. Bu bulgu ışığında, Krippendorff alfa tekniği ile uyum elde edilmek istendiğinde iki ile beş arasında puanlayıcıya yer vermenin yeterli olduğu; beş puanlayıcıdan sonra istatistikte çok değişim olmadığı söylenebilir.

Çalışmada, analitik dereceli puanlama anahtarına dayalı olan log linear analiz sonuçlarına göre puanlayıcılar arasındaki uyumun yanı sıra puanlayıcıların, kategorilerin ve altkategorilerin birbirleriyle etkileşimi de elde edilmiştir. Bu da araştırmacıya daha ayrıntılı bilgi sunmuş ve diğer tekniklerden farklı olarak sonuçlardaki uyumsuzluğun nereden kaynaklandığı konusunda bilgi vermiştir. Analiz sonuçlarının uyumsuzluk hakkında bilgi vermesi puanlayıcılar arasındaki uyumsuzluğa neden olan değişkenlerin araştırmadan çıkarılmasına imkân sağlayabilmektedir.

Araştırma neticesinde, daha detaylı ölçme sonuçları elde edilmek istendiğinde alt kategorilerden oluşan analitik puanlama anahtarı kullanılarak toplanan puanlar, kategorik veri analizi için uygun olan log linear analiz tekniği ile analiz edilebilir. Diğer açıdan daha genel ölçme sonuçlarına ulaşılmak istendiğinde bütünsel puanlama anahtarı ile elde edilen puanların Krippendorff alfa tekniği kullanılarak analiz edilmesinin uygun olduğu düşünülebilir. Sonuç olarak, performansın ölçülmesinde güvenirliğin belirlenmesinde yararlanılacak tekniklerin hangisinin seçileceği, elde edilen puanların hangi amaç doğrultusunda kullanılacağına, puanların hangi ölçek türünde elde edildiğine, analiz sonucunda sağladıkları bilgilerin yapılan ölçme işleminin amacına uygunluğuna bağlı olarak değişmektedir.

KAYNAKÇA

Abedi, J., Baker, E L., & Herl, H. (1995). Comparing reliability indices obtained by different approaches for performance assessments. Los Angeles: University of California, *CSE Technical Report, 401*.

Airasian, P. W. (1994). *Classroom assessment.* New York: McGraw-Hill.

Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons, INC.

Agresti, A. & Yang, M. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis, 5*, 9-21.

Atılgan, H., Kan, A. ve Doğan, N. (2007). *Eğitimde ölçme ve değerlendirme* (2. Basım). Ankara: Anı.

Baykul, Y. (2000). *Eğitim ve psikolojide ölçme: Klasik Test Teorisi ve uygulaması.* Ankara: ÖSYM.

Brennen, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*(1981), 687-699.

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater-agreement indexies for performance assessment. *Educational and Psychological Measurement, 56*(2), 251-262.

Cohen. J. R., Swerdlik M. E., & Phillips, S. M. (1996). *Psychological testing and assessment*. (3th Ed.). London: Mayfield.

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

Crawforth, K. (2001). M*easuring the interrater reliability of a data collection instrument developed to evaluate anesthetic outcomes* (Doctoral Dissertation). Available from Proquest Dissertations and Theses database. (UMI No. 3037063)

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Ohio: Centage Learning.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (p. 237–270). Washington DC: American Council on Education.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Goodrich, H. (1997). Understanding rubric. *Educational Leadership, 54*(4), 14-17.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Psychical Education and Exercises Science*, *5*(1), 13-14.

Haladyna, M. T. (1997). *Writing test items to evaluate higher order thinking*. Needham Heights: Allyn and Bacon.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Rewiew*, *2*(2007), 130-144.

Korkmaz, H. (2004). *Fen ve teknoloji eğitiminde alternatif değerlendirme yaklaşımları.* Ankara: Yeryüzü.

Krippendorff, K. (1995). On the reliability of unitizing continuous data. *Sociological Methodology, 25,* 47-76.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Humanities, Social Sciences and Law, 38*(6), 787-800.

Krippendorff, K. (2007). Computing Krippendorff's alpha reliability. 7 Eylül 2015 tarihinde http://repository.upenn.edu/asc_papers/43/ adresinden erişildi.

Kutlu, Ö., Doğan, D. C. ve Karakaya, İ. (2009). *Öğrenci başarısının belirlenmesi: performansa ve portfolyaya dayalı durum belirleme.* Ankara: Pegem Akademi.

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

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Pratical Assessment Research and Evaluation, 7*(25). Available online: http://PAREonline.net/getvn.asp?v=7&n=25.

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment Research and Evaluation, 7*(3). Available online: http://PAREonline.net/getvn.asp?v=7&n=3.

Nitko, A. J. (2001). *Educational assessment of students.* (3th ed). New Jersey*:* Prentice Hall.

Nying, E. (2004). *A comparative study of interrater reliability coefficients obtained from different statistical procedures using monte carlo simulation tecniques* (Doctoral Dissertation). Available from Proquest Dissertations and Theses database. (UMI No. 3138768).

Sim, J., & Wright, C. C. (2005) The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Theraphy*, *85*(3), 258-268.

Tanner, M. A., & Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, *80*(389). 175-180.

Viere, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360-362.

Von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement*: *Manifest variable methods*. New Jersey: Lawrence Erlbaum Associates.

EXTENDED ABSTRACT

Introduction

The concept of performance based assessment has gained currency in recent years instead of traditional conception of measurement. Analytic and holistic rubrics are used in performance based assessments as the instruments of measurement. The instruments employed in performance assessment should be valid and reliable, as in other instruments of measurement. Reliability is defined as freedom of measurement results from random errors (Baykul, 2000). In the process of performance-based assessment, where raters are the undesired source of variability, techniques for interrater reliability assessment have been recommended. Interrater reliability is defined as no change of rating from one rater to another (Kutlu et al., 2009). It is apparent from the literature that several techniques have been

employed in evaluating interrater reliability (Jonsson & Svingby, 2007). This study aims to analyse and compare the reliability values calculated for the performance prepared for the same purpose and administered to the same individuals by using analytic and holistic rubric through Kappa, Krippendoff, Log-linear analysis techniques. It also aims to find whether rater reliability varies depending on the number of raters and the techniques used.

Method

This study is built on the basis of applying Kappa statistis, Krippendorff Alpha coefficient and log-linear analysis techniques, on determining the similarities as well as differences of these techniques, examining their restrictions, and on finding which of these techniques provide more information. Since it tries to determine the situation, it is a descriptive study.

The study group was composed of 50 students fulfilling performance tasks and reporting them, and 10 teachers rating the tasks by using two rubrics: one of the rubrics was analytic and the other one was holistic. The students in the study group were 50 students who were the 5^{th} graders in a state school located in Beyoglu district of Istanbul- where one of the researchers was teaching. Such factors as reaching the students easily and being able to monitor them during application were influential in determining the participants. The teachers in the study group were the elementary school teachers employed in Istanbul in the body of the Ministry of National Education. The participating teachers were chosen on the basis of volunteering.

Analytic and holistic rubrics were used in this study as the tool of data collection. Analytic rubric, developed by Kutlu et al (2009), was composed of five sub-categories-namely, content, the process of research, use of materials, graphic formation, and tabulation. In each sub-category performance was scored between 1 and 3. The maximum score receivable was 15 while the minimum score receivable was 5. Holistic rubric was created on the basis of analytic rubric, which was developed by Kutlu et al (2009) and its final shape was given by consulting five measurement and evaluation experts' opinions. The maximum score receivable from this instrument was 4 whereas the minimum score was 1.

Results and Discussion

Ratings obtained by using analytic rubric yielded relatively more objective results than the one obtained by using holistic rubric- even though both rubrics were designed for the same purpose. Thus, it was concluded that analytic rubric- which produced more standard and more objective results- yielded more consistent results and that therefore it was more reliable. This finding is in parallel to the one obtained by Kutlu et al (2009) stating that the results obtained with the use of holistic rubric were less reliable than those obtained with the use of analytic rubric. The finding is also supportive of Johnsson and Svingby's (2007) conclusion that analytic rubrics increase reliability.

This study analysed the results obtained by using analytic rubric containing sub-categories. It was found that the categories constituting the rubric raised the level of objectivity in the order of content, the process of research, use of materials, graphic formation and tabulation. Based on this finding, it may be said that using analytic rubric is not adequate on its own in eliminating interrater differences completely but that it increases interrater consistency and thus it raises the level of objectivity.

The highest reliability values in analyses performed through all three techniques in the scores obtained from both analytic and holistic rubrics were reached when there were two raters, and it was observed that reliability decreased gradually as the number of raters increased. This finding is in parallel to the Abadi, Baker, Herl's (1995) finding that increase in the number of raters causes a rise in the level of variability in scores and a decrease in reliability in performance measurement, and to Nying's (2004) finding that reliability is influenced by increase in the number of raters.

The highest agreement between raters was observed when there were two raters in all three analyses performed by using Kappa statistics and Krippendorff alpha techniques. The increase in the number of raters reduced the value; however, the significance level of the statistics remained the same. On

examining the Kappa and Alpha results together, it was found that the increase in the number of raters brought about a change in values but that the change varied as much in Alpha results as in Kappa statistics and that it was stable. On raising the number of raters from two to five in analyses, a fall was observed in Alpha values, but when the number of raters was raised to ten, no significant decrease was seen in Alpha values.

The interactions of raters, categories, and sub-categories in addition to interrater agreement were also obtained in log linear analysis results in this study. This provided the researcher with more detailed information, and as different from other techniques, it also informed the researcher about the source of disagreement between analysis results. Presenting information on the disagreement between analysis results enables researchers to remove the variables causing disagreement from their study.

If the intention is to obtain more detailed measurement results from research, scores obtained by using the analytic rubric which is composed of sub-categories can be analysed by using log-linear analysis technique- which is suitable for categorical data analysis. On the other hand, if the purpose is to obtain more general measurement results, analysing the scores obtained through holistic rubric by using the Krippendorff technique would be more appropriate. In consequence, which technique to choose in determining reliability in performance measurement changes depending on for what purpose to use the scores obtained, in what type of scale the scores have been obtained, and on the suitability of the data provided in consequence of the analyses to measurement purpose.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);79-97

Madde Tepki Kuramı'na Dayalı Madde-Uyum İndekslerinin I.Tip Hata ve Güç Oranlarının İncelenmesi*

Investigating Type I Error and Power Rates of Item Fit Indices Based on Item Response Theory

Seçil ÖMÜR SÜNBÜL** Semih AŞİRET***

Öz

Bu çalışmada, Madde Tepki Kuramı'na göre ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin, çeşitli koşullardaki (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi) I. tip hata ve güç oranlarının incelenmesi amaçlanmıştır. Çalışmada, indekslerin I. tip hata ve güç oranlarının belirlenmesi simülasyon çalışmasıyla yapılmıştır. Çalışmada, madde uyumu için geleneksel indekslerden χ^2, Q_1 ve G^2 indeksleri ile alternatif indekslerden $S\text{-}\chi^2$ indeksi kullanılmıştır. Çalışmada yer alan dört farklı madde-uyum indeksinin I. tip hata ve güç oranları, örneklem büyüklüğü (1000, 2000, 4000), test uzunluğu (20, 30, 40) ve uyumsuzluk yüzdesi (%0, %10, %30 ve %50) değişimlenerek incelenmiştir. Veriler R 3.3.2 yazılımı kullanılarak üretilmiştir ve "mirt" paketi kullanılarak analiz edilmiştir. Çalışmada üretilen ve analiz edilen model olmak üzere iki tür model kullanılmıştır. Üretilen modele uygun madde tepkileri ile analiz edilen modele uygun madde tepkileri için madde-uyum indekslerinin *p* değerleri ve serbestlik dereceleri hesaplanmıştır. Uyum indekslerinin I. tip hata ve güç oranları 0.05 anlamlılık düzeyine göre değerlendirilmiştir. Her uyum indeksinin tüm koşullardaki I. tip hata ve güç oranları hesaplanarak bu indeksler karşılaştırılmıştır. Çalışma sonucunda, tüm faktörlerde $S\text{-}\chi^2$ indeksinin diğer indekslere göre daha düşük hataya sahip olduğu görülmüştür. 2000 ve üzeri örneklem büyüklüğünde ve 20 ve daha fazla maddeden oluşan testlerde $S\text{-}\chi^2$ indeksinin diğer indekslerden daha düşük I. tip hata oranına ve daha yüksek güce sahip olduğu görülmüştür.

Anahtar Kelimeler: Madde Tepki Kuramı, madde-uyum indeksi I.tip hata, güç, $S\text{-}\chi^2$

Abstract

In this study, it was aimed to investigate type I error and power rates of the item fit indices through various conditions (sample sizes, different test lengths and different magnitudes of misfit) for dichotomously generated items based on one-, two-, and three-parameter logistic models in Item Response Theory. In this study, the type I error and power rates of these item fit indices were assessed in a simulation study. χ^2, Q_1 and G^2 indices as traditional item fit indices and $S\text{-}\chi^2$ index as alternative indices were assessed. The performance of four different item fit indices in study were compared by manipulating three different sample size (1000, 2000, 4000), three different test lengths (20, 30, 40) and four different misfit magnitude (%0, %10, %30 and %50). Item responses were generated using the R 3.3.2 software program and analyzed by using "mirt" package in R software. The p value of item fit indices and their degrees of freedom were calculated for both item responses for generating model and analysis model. Type I errors and power rates of item fit indices were examined according to significance levels of 0.05. All item fit indices in this study were compared by calculating the type I error and power rates of each item fit indices under all conditions. The findings indicated that $S\text{-}\chi^2$ index has lower type I error to detect misfit than the other indices. It can be concluded that in the case where the sample size was 2000 or more and the number of items in test are 20 and more, $S\text{-}\chi^2$ index has lower type I error rates than traditional indices and has adequate power to detect misfit items.

Keywords: Item Response Theory, item fit index, type I error, power, $S\text{-}\chi^2$.

* Bu çalışma V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde (01-03 Eylül 2016) sözlü bildiri olarak sunulmuştur.

** Yrd. Doç. Dr., Mersin Üniversitesi, Eğitim Bilimleri Bölümü, Mersin-Türkiye, e-posta:secilomur@gmail.com
*** Uzman, Mersin Üniversitesi, Eğitim Bilimleri Bölümü, Mersin-Türkiye, e-posta:semihasiret@gmail.com

DOI: 10.21031/epod.301529
Geliş Tarihi: 15.11.2016
Kabul Tarihi: 13.03.2017

GİRİŞ

Madde Tepki Kuramı (MTK), madde ve bireyin özelliklerini kullanarak, bireyin performansını kestirmede matematiksel modeller kullanan güçlü bir ölçekleme tekniğidir (Embretson ve Reise, 2000). Madde Tepki Kuramı'nın test puanlarını yorumlamada ve test sonuçlarını raporlamada birçok avantajı bulunmaktadır. Ancak, bu avantajlar seçilen model ile test verilerinin uyumlu olduğu durumlarda elde edilebilir (Hambleton ve Swaminathan, 1985). Madde Tepki Kuramı'nın güçlü yanlarından bir tanesi de değişmezlik özelliğine dayalı olmasıdır, yani madde parametrelerinin yetenek dağılımlarına veya birey parametrelerine bakılmaksızın aynı kalmasıdır (Embretson ve Reise, 2000). Ancak parametre değişmezliği, belirli MTK modellerinin tek boyutluluk, yerel bağımsızlık ve madde karakteristik eğrisinin monotonik artması gibi sayıltıları gerçekleştikten sonra geçerli olacaktır (Wells ve Hambleton, 2016). MTK modeli veriyle uyumsuzken uygulandığında değişmezlik özelliğindeki tüm önemli bilgiler kaybolacaktır. MTK modellerinin geçerli bir şekilde uygulanması ve kararlı puan ölçeklerinin elde edilebilmesinde, model veri uyumu önemli bir role sahiptir.

Aynı modelin testteki tüm maddelere uygulanma zorunluluğu yoktur. Örneğin; bir test, hem ikili hem de çoklu puanlanan maddelerden oluşabildiği gibi, aynı zamanda bazı maddeler iki parametreli lojistik model ile bazıları ise aşamalı tepki (graded response) modeliyle uyum gösterebilir (Embretson ve Reise, 2000). Bu nedenle, birçok çalışmada genel model-veri uyumunun aksine, MTK model uyumunun madde madde yargılanması önerilmektedir (Chon, Lee ve Ansley, 2007; Chon, Lee ve Dunbar, 2010; Tay, Ali, Drasgow ve Williams, 2011; Wells ve Bolt, 2008).

Belirli bir madde düzeyinde, maddelerin uyumunun değerlendirilmesinde genel strateji, gözlenen verilerle kestirilen verinin karşılaştırılmasıdır (Hambleton, Swaminathan ve Rogers, 1991). MTK'de öncelikle MTK modelinin parametreleri kestirilir. Ardından bu kestirilen parametreler kullanılarak bireylerin tepki örüntüleri kestirilir. Son olarak kestirilen tepki örüntüleri ile bireyin gerçek gözlenen tepki örüntüleri karşılaştırılır.

Madde uyumunu değerlendirme işlemi iki genel yaklaşımla gerçekleştirilir. Birincisi, çok fazla istatiksel işlem gerektirmeyen grafiksel işlemlerdir. Burada madde uyumunun yargılanması, kestirilen Madde Tepki Eğrisi (MTE) ile görgül olarak gerçek veya gözlenen verilerden elde edilen MTE'nin karşılaştırılmasına dayanır (Embretson ve Reise, 2000; Reise, 1990). Kestirilen MTE ile gözlenen MTE arasındaki fark artık (residual) olarak adlandırılır ($r=O_{ig}-E_{ig}$ şeklinde gösterilir). Artıkların görsel olarak gösterimiyle model veri uyumunun incelenmesi faydalı görülürken, öznel olma durumundan dolayı da eleştirilmiştir. Bu sebeple, MTK modellerinin veri ile uyumunu incelemek için birçok madde-uyum indeksi geliştirilmiştir (Ames, 2015; Ames ve Penfield, 2015; Lahuis, Clark ve O'brien, 2011).

Birçok madde-uyum indeksi geliştirilmiş olsa da, bu indeksler genel olarak ki-kare yaklaşımı ve olabilirlik oranı yaklaşımı olmak üzere iki yaklaşıma dayanmaktadır. İkili puanlanan maddelerde model veri uyumunu değerlendirmedeki ki-kare yaklaşımı genel olarak Eşitlik-1 de gösterilmektedir.

$$\chi^2 = \sum_{g=1}^{G} N_g \frac{(O_{ig}-E_{ig})^2}{E_{ig}(1-E_{ig})} \quad (1)$$

Eşitlik-1'de, O_{ig}, g aralığındaki i maddesi için gözlenen doğru oranını, E_{ig}, yetenek kestirim aralığındaki kestirilen MTE'ye dayalı beklenen doğru oranını, N_g, g yetenek aralığına denk gelen birey sayısını göstermektedir. Ki-kare istatistiği standardize edilmiş artıkların karesinin toplamını göstermektedir. Artık, doğrudan ki-kare eşitliğinin içinde yer almaktadır. Artık değerleri arttıkça ki-kare değeri de artmaktadır.

İkili puanlanan maddeler için olabilirlik oranı ise Eşitlik-2'de gösterilmektedir.

$$2\sum_{k=1}^{K} N_k\left[O_{ik}\ln\left(\frac{O_{ik}}{E_{ik}}\right)+(1-O_{ik})\ln\left(\frac{1-O_{ik}}{1-E_{ik}}\right)\right] \quad (2)$$

Eşitlik-2'de ilk bakışta artıklar görülmez. Ancak $\ln(\frac{O_{ik}}{E_{ik}})$'nin doğal logaritması aynı zamanda $\ln O_{ik} - \ln E_{ik}$ şeklinde ifade edildiğinden eşitlikte artıklar yer almaktadır.

Ki-kare veya olabilirlik oranı yaklaşımlarından bir tanesine dayalı olan madde uyum indeksleri iki farklı boyutta farklılaşır (Ames ve Penfield, 2015). Birincisi, grupların oluşturulma şeklidir. Gruplar yetenek düzeylerine göre farklı şekillerde oluşturulabilmektedir. İkincisi ise, kestirilen madde tepki eğrisindeki doğru cevaplama oranlarının (E_{ig}) hesaplanma şeklidir.

Lahuis, Clark ve O'brien (2011) bu indekslerin serbestlik derecelerinin hesaplanma şekline bağlı olarak farklı şekilde sınıflandırılabileceğini belirtmiştir. Buna göre indeksler, geleneksel ve alternatif madde-uyum indeksleri olmak üzere iki şekilde sınıflandırılmıştır. Geleneksel madde-uyum indekslerinde model-veri uyumu, her maddenin seçilen MTK modelindeki çeşitli alt yetenek gruplarının gözlenen performansı ile kestirilen performansı karşılaştırılarak değerlendirilir (Stone ve Zhang, 2003). Geleneksel madde-uyum indeksleri; *OUTFIT* ve *INFIT* indeksleri (Wright ve Panchapakesan, 1969), Bock'un X^2 indeksi (Bock, 1972), Yen'in Q_1 istatistiği (Yen, 1981) ve G^2 indeksidir (McKinley ve Mills, 1985).

Geleneksel yöntemlerde yetenek aralıklarının belirlenmesi genellikle keyfidir ve belirlenen farklı aralıklar oluşabilecek sonuçları da etkilemektedir (Orlando ve Thissen, 2000; Reise, 1990). Aralıklar yetenek koşuluna bağlı olduğu için gözlenen tepkiler modele bağımlı olmaktadır. Bu durum, uyum indekslerinin serbestlik derecesini etkileyebilmektedir (Orlando ve Thissen, 2000). Stone ve Zhang (2003), yetenek kestirimindeki belirsizliğin χ^2 istatistiği üzerinde olumsuz bir etkiye sahip olduğunu belirtmiştir. Ayrıca geleneksel indeksler test uzunluğu ve örneklem büyüklüğüne karşı çok duyarlıdır (Kang ve Chen, 2011). Literatürde geleneksel madde-uyum indekslerin sıklıkla kullanıldığı görülmekle birlikte bu indekslerin sınırlılıklarıyla ilgili detaylı birçok çalışmanın yapıldığı da görülmektedir (DeMars, 2005; Glas ve Su´arez Falc´on, 2003; Orlando ve Thiseen, 2000; Stone ve Zhang, 2003; von Schrader, Ansley ve Kim, 2004). Bu sebeple, birçok alternatif madde-uyum indeksleri üretilmiştir (Lahuis, Clark ve O'brien, 2011). Bu indeksler Orlando ve Thissen (2000) tarafından geliştirilen *S*-χ^2, Stone (2000) tarafından gerçekleştirilen ölçeklenmiş düzeltilmiş (scaling corrected) uyum istatistiği (χ^{2*}) ve Drasgow, Levine, Tsien, Williams ve Mead (1995) tarafından geliştirilen ayarlanmış (adjusted) χ^2-serbestlik dereceleri oranı (χ^2/dfs) indeksidir. Bu çalışma kapsamında kullanılacak yazılım farklılığından kaynaklanacak hataların oluşmaması açısından tüm madde uyum indekslerinin aynı yazılımla analiz edilmesi amaçlanmış ve bu nedenle geleneksel madde uyum indekslerinden X^2, Q_1, G^2 ve alternatif madde uyum indekslerinden *S*- X^2 kullanılmıştır. Aşağıda bu çalışma kapsamında kullanılan indekslerden kısaca bahsedilmiştir.

Bock'un Ki-Kare İndeksi (Bock, 1972)

Bock'un ki-kare eşitliği Eşitlik-3 kullanılarak elde edilmektedir. Eşitlik-3'te O_{ig}, g aralığındaki i maddesi için gözlenen doğru oranını, E_{ig}, yetenek kestirim aralığındaki ortancada kestirilen MTE'ye dayalı beklenen doğru oranını, N_g, g yetenek aralığına denk gelen birey sayısını göstermektedir. Ki-kare dağılımının serbestlik derecesi (*Gxm*) ile elde edilir (*G*, aralık sayısı ve m, madde sayısıdır) (Embretson ve Reise, 2000).

$$BCHI = \sum_{g=1}^{G} \frac{N_g(O_{ig} - E_{ig})^2}{E_{ig}(1 - E_{ig})} \qquad (3)$$

Bock'un ki-kare istatistiği beklenen frekansları kestirmek için yetenek aralıklarının ortancasını kullanır ve bu aralıkların boyutları birbirinden farklıdır (Lahuis, Clark ve O'brien, 2011).

Yen'in Q_1 İndeksi

Yen'in Q_1 indeksi (1981), yeteneği 10 eşit aralığa böler ve beklenen frekansları hesaplamak için aralıkların ortalamasını kullanır (Lahuis, Clark ve O'brien, 2011).Yen'in Q_1 istatistiği Eşitlik-4'te belirtilen şekilde hesaplanır.

$$Q_1 = \sum_{j=1}^{10} N_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})} \quad (4)$$

Yen'in Q_1 indeksi ve Bock'un χ^2 indeksi ile yapılan genel eleştirilerden bir tanesi bu iki istatistiğin grupları oluştururken yetenek değerlerini kullanmalarıdır. Model uyumu zayıf olduğunda, kestirilen yetenek değerleri yanlı olacaktır. Yanlı yetenek kestirimlerine göre oluşturulan gruplardan elde edilen istatistikler ise yanlış madde-uyum istatistiği verebilir. Bu yöntemlerin diğer bir sınırlılığı ise, gruplar oluşturulurken keyfi kesme puanı belirlenerek her gruba bireylerin yerleştirilmesidir. Bu durumda, bu istatistikler gruba dayalı istatistikler olacaktır.

Bu indeksler, mükemmel model veri uyumunun sıfır hipotezini değerlendirirler. İndekslerden elde edilen değerler ile serbestlik derecelerine denk gelen kritik değerler karşılaştırılarak hipotez red veya kabul edilir.

G^2 İndeksi

McKinley ve Mills (1985), tarafından geliştirilen bu indeks χ^2 olabilirlik oranı olarak adlandırılır. G^2 indeksi Yen'in Q_1 indeksine benzemektedir. Bu indekste, yetenek 10 eşit aralığa bölünür ve gözlenen ve beklenen frekanslar karşılaştırılır (Lahuis, Clark ve O'brien, 2011). Serbestlik derecesi grup sayısına eşittir. G^2 istatistiği Eşitlik-5'te gösterilen şekilde elde edilmektedir.

$$G_i^2 = 2\sum_{k=1}^{10} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1 - O_{ik}) \ln\left(\frac{1 - O_{ik}}{1 - E_{ik}}\right)\right] \quad (5)$$

Bu indeks BILOG-MG ve PARSCALE yazılımlarında standart model veri indeksi olarak kullanılmaktadır. Bock'un χ^2 ve Q_1 istatistikleri gibi bu istatistikte grupların yetenek değerlerine bağlı olarak oluşturulmaktadır.

S-χ^2 İndeksi

Orlando ve Thissen (2000), tarafından ikili puanlanan maddeler için geliştirilen madde-uyum indeksidir. Bu indeks yetenek yerine toplam puanlar üzerine koşullanmıştır. Toplam puandan elde edilen gözlenen ve beklenen tepkiler, χ^2 istatistiği kullanılarak karşılaştırılır. Beklenen tepkilerin hesaplanmasında, her toplam puan için, tüm olası tepki örüntülerinin, her olası toplam puanının katışık olabilirlik dağılımı kullanılır. Yani beklenen tepkiler verilen maddenin başarıldığı ve toplam puanın üretildiği tepki örüntülerinin olabilirliğine dayalıdır. Beklenen tepkiler Thissen, Pommerich, Billeaud ve Williams (1995) tarafından geliştirilen özyinemeli algoritmalar kullanılarak geliştirilir (Lahuis, Clark ve O'brien, 2011). Beklenen tepkiler Eşitlik-6'da gösterilen şekilde hesaplanır.

$$E_{ik} = \frac{\int P_{i1}(\theta) f^{*i}(k-1|\theta) \phi(\theta) d\theta}{\int f(k|\theta) \phi(\theta) d\theta} \quad (6)$$

Eşitlik-6'da P_{i1}, i maddesi için tepki fonksiyonunu, $f(k|\theta)$; verilen yetenekteki koşullu kestirilen test puan dağılımını, $f^{*i}(k-1|\theta)$; i maddesi olmadan koşullu kestirilen test puan dağılımını ve $\phi(\theta)$; yetenek dağılım evrenini göstermektedir. Orlando ve Thissen (2000), χ^2 ve G^2 indekslerini karşılaştırarak ki-kare indeksini (S-χ^2) ve olabilirlik oran istatistiğini (S-G^2) hesaplamıştır. Bu

istatistiklerin serbestlik derecesi toplam kategori sayısından (maksimum elde edilecek puan -1) madde parametre sayısının çıkartılmasıyla elde edilir. Gerektiğinde hücrelerde minimum beklenen frekansın (1) elde edilmesi için hücreler daraltılır. Daralma olduğunda serbestlik derecesinde de düzenlemeler yapılır.

(S-χ^2) indeksinin birçok olumlu özelliği bulunmaktadır. Bu indeksin ikili ve çoklu puanlanan MTK modelleri için I. tip hatası kabul edilebilir ve bu indeks büyük örneklemlerde (N≥2000) yeterli güce sahiptir (Kang ve Chen, 2008; Orlando ve Thissen, 2000, 2003; Stone ve Zhang, 2003). Ayrıca özyinelemeli algoritmalar, ikili puanlanan maddeler için GOODFIT yazılımı (Orlando, 1997), R yazılımı ve hem ikili hem de çoklu puanlanan maddeler için IRTFIT yazılımı ile uygulanabilmektedir. (S-χ^2) indeksi yeteneği keyfi aralıklara bölmez çünkü toplam puan koşuluna bağlıdır. Ancak madde sayısı veya tepki seçenekleri arttığında olası toplam puan sayısı da artacaktır. Bu durum sorun oluşturabilmektedir. Ayrıca bu durumda indeksi hesaplamak büyük emek isteyebilir.

Bu çalışmada MTK'ye dayalı bir, iki ve üç parametreli lojistik modeller için farklı koşullarda madde-uyum indekslerinin I. tip hata ve güç oranlarının değerlendirilmesi amaçlanmıştır. İlgili literatür incelendiğinde, yapılan çalışmalarda genellikle örneklem büyüklüğü ve test uzunluğu faktörlerinin madde-uyum indeksleri üzerindeki etkilerinin incelendiği görülmüştür (Chon, Lee ve Ansley, 2007; Chon, Lee ve Dunbar, 2007; DeMars, 2005; Glas ve Su´arez Falc´on, 2003; Orlando ve Thissen, 2000, 2003; Reise, 1990; Stone ve Zhang, 2003; Wells ve Bolt, 2008). Test uzunluğu ile birlikte testte yer alan uyumsuz madde miktarının da madde-uyum indekslerinin performanslarına etki edeceği düşünülmektedir. İlgili literatürde uyumsuz (misfit) madde miktarının madde-uyum indekslerinin I. tip hata ve güç oranları ile ilgili yeterli çalışma olmadığı görülmüştür (Wells ve Bolt, 2008; Ames, 2015). Bununla birlikte ilgili faktörlerin ortak etkisini inceleyen bir çalışmaya da rastlanılmamıştır. Ayrıca madde-uyum indeksleriyle ilgili ülkemizde yapılan ilk araştırma olması nedeniyle alana katkı getireceği düşünülmektedir.

Çalışma kapsamında aşağıdaki sorulara cevap aranmaya çalışılmıştır.

1. Çeşitli faktörlerin (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi), madde-uyum indekslerinin (X^2, Q_1, G^2 ve S-X^2) I. tip hata ve güç oranlarına temel etkisi nasıldır?
2. Çeşitli faktörlerin (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi), madde-uyum indekslerinin (X^2, Q_1, G^2 ve S-X^2) I. tip hata ve güç oranlarına ortak etkisi nasıldır?

Araştırmanın Amacı

Bu çalışmada, Madde Tepki Kuramı'na göre, ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin çeşitli koşullardaki (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi) I. tip hata ve güç oranlarının incelenmesi ve hangi koşullarda hangi indeksin daha iyi sonuç verdiğinin belirlenmesi amaçlanmıştır.

YÖNTEM

Aşağıdaki bölümde araştırmanın türünden, bu araştırma kapsamında değişimlenen faktörlerden ve bunların düzeylerinden, verilerin üretiminden ve işlem adımlarından bahsedilmiştir.

Araştırmanın Türü

Bu çalışmada, Madde Tepki Kuramı'na göre, ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin çeşitli koşullardaki I. tip hata ve güç oranlarının incelenmesi ve hangi koşullarda hangi indeksin daha iyi sonuç verdiğinin belirlenmesi amaçlandığından çalışma, temel araştırma olarak değerlendirilebilir.

Araştırma Kapsamında Değişimlenen Faktörler

Bu çalışmada, indekslerin I. tip hata ve güç oranlarının belirlenmesi simülasyon çalışmasıyla yapılmıştır. Çalışma kapsamında, madde uyumu için geleneksel indekslerden χ^2, Q_1 ve G^2 indeksleri ile alternatif indekslerden (S-χ^2) indeksi kullanılmıştır. Çalışmada, örneklem büyüklüğü (1000, 2000, 4000), test uzunluğu (20, 30, 40) ve uyumsuzluk yüzdesi (%0, %10, %30 ve %50) faktörleri değişimlenmiştir. Çalışmada yer alan faktörler ve düzeyleri Tablo-1'de özetlenmiştir.

Veri Üretimi

Bu çalışma kapsamında veriler, üretilen model (GM) ve analiz edilen model (CM) olmak üzere iki farklı şekilde, değişimlenen faktörlere uygun olarak üretilmiştir. Veri üretiminde, MTK'ye dayalı 2PL ve 3PL modelleri kullanılmıştır. İlk olarak GM için, kullanılacak modele (2PL ya da 3PL), değişimlenen faktöre ve bunların düzeylerine uygun olarak veri üretilmiştir. Üretilen 2PL ve 3PL modeller için a parametre değerleri 1.00, c parametre değerleri 0.00 ve b parametre değerleri minimum -2 maksimum +2 olan uniform dağılımdan elde edilecek şekilde ayarlanmıştır. Daha sonra analiz edilen model için (CM), veri üretiminde kullanılan modele uygun olarak a, b ve c parametreleri değişimlenerek veri üretilmiştir. Analiz edilen 1PL, 2PL ve 3PL modeller için b parametre değerleri üretilen modeldeki b parametre değerlerine $\pm$ 0.75 eklenerek değişimlenmiştir. 2PL ve 3PL modeller için a parametre değerleri minimum 0, maksimum +2 olan uniform dağılımdan çekilerek değişimlenmiştir. 3PL model için ise c parametre değerleri ise 0.25 olacak şekilde değişimlenmiştir. Çalışmada bireylerin yetenek dağılımlarına ilişkin değerleri ise; ortalaması 0, standart sapması 1 olan standart normal dağılımdan *N(0,1)* elde edilmiştir. Bu yetenek ve parametre değerlerine göre 1-0 verileri üretilmiştir.

Tablo 1. Çalışmada Değişimlenen Faktörler ve Düzeyleri

Faktör	Düzey Sayısı	Düzey Değerleri
Örneklem Büyüklüğü	3	1000 2000 4000
Madde Sayısı	3	20 30 40
Uyumsuzluk Yüzdesi	4	%0 %10 %30 %50
Replikasyon Sayısı	100	

Verilerin Analizi

Verilerin üretiminden sonra ilk olarak maddelerin uyumsuzluk miktarları hesaplanmıştır. Madde modele uyumsuz olsa da uyumsuzluğun miktarı küçük olduğunda uyum indeksleri uyumsuz maddeleri tespit etmede zorlanabilmektedir. Bu sebeple, Wells ve Bolt (2008), tarafından geliştirilen *MISFIT* indeksi kullanılarak uyumsuz maddelerin uyumsuzluk miktarları hesaplanmıştır. Wells ve Bolt (2008), 0.020 ve üzeri uyumsuzluk miktarlarını orta ve yüksek uyumsuzluk olarak ele alınabileceğini ifade etmiştir. Bu sebeple çalışmada uyumsuz olarak kalibre edilmiş maddelerin uyumsuzluk miktarları 0.020 ve üzeri olanlar tercih edilmiştir. GM=3PL ve CM= 2PL, CM=1PL durumlar için üretilen a, b ve c parametreleri ile uyumsuzluk büyüklükleri Tablo 2'de verilmiştir.

Analiz edilen her bir model için maddelerdeki uyumsuzluk miktarı aşağıda belirtilen Eşitlik-7 ile hesaplanmıştır.

$$MISFIT = \sqrt{\sum_{j=1}^{601} w(\theta_j)\,(P_{GMj} - P_{CMj})^2} \qquad (7)$$

MISFIT indeksi, üretilen ve analiz edilen modelin tepki olasılıkları farkının karelerinin ağırlıklandırılmasıyla elde edilen sonucun toplamına eşittir. Yetenek -3 ile + 3 arasında 601 eşit parçaya bölünmüştür. $j=1,2,\ldots\theta$ için $w(\theta_j)$, standart normal yoğunluk tarafından tanımlanan normalleştirilmiş ağırlıktır. P_{GMj} ise θ_j yetenek düzeyindeki bireyin üretilen model için maddeyi cevaplandırma olasılığı iken, P_{CMj} , θ_j yetenek düzeyindeki bireyin analiz edilen model için maddeyi cevaplandırma olasılığıdır. Uyumsuzluk miktarı 0.020 ve üzeri olan maddeler orta ve yüksek uyumsuzluk miktarı olarak ele alınmaktadır.

Tablo 2. 3PL Model İçin Üretilen Parametre Değerleri ve Uyumsuzluk İndeksi Değerleri

	Uyumsuzluk (%0)			Uyumsuzluk (%10)			Uyumsuzluk indeksi		Uyumsuzluk (%30)			Uyumsuzluk indeksi		Uyumsuzluk (%50)			Uyumsuzluk indeksi	
M.N	a	b	c	a	b	c	2PL	1PL	a	b	c	2PL	1PL	a	b	c	2PL	1PL
1	1.00	-1.96	0.00	1.00	-1.96	0.00	-	-	**0.47**	**-1.96**	**0.25**	**0.026**	**0.034**	**0.41**	**-1.24**	**0.25**	**0.027**	**0.038**
2	1.00	-1.84	0.00	1.00	-1.68	0.00	-	-	1.00	-1.84	0.00	-	-	1.00	-1.90	0.00	-	-
3	1.00	-1.69	0.00	1.00	-1.48	0.00	-	-	1.00	-1.69	0.00	-	-	**0.57**	**-1.00**	**0.25**	**0.028**	**0.033**
4	1.00	-1.63	0.00	1.00	-1.41	0.00	-	-	1.00	-1.63	0.00	-	-	1.00	-1.73	0.00	-	-
5	1.00	-1.37	0.00	**0.16**	**-0.59**	**0.25**	**0.037**	**0.070**	**0.16**	**-0.59**	**0.25**	**0.039**	**0.085**	**0.16**	**-0.59**	**0.25**	**0.034**	**0.064**
6	1.00	-1.31	0.00	1.00	-1.02	0.00	-	-	1.00	-1.31	0.00	-	-	1.00	-1.72	0.00	-	-
7	1.00	-1.06	0.00	1.00	-0.96	0.00	-	-	1.00	-1.06	0.00	-	-	**0.47**	**-0.97**	**0.25**	**0.029**	**0.039**
8	1.00	-0.93	0.00	1.00	-0.93	0.00	-	-	1.00	-0.93	0.00	-	-	1.00	-1.48	0.00	-	-
9	1.00	-0.81	0.00	1.00	-0.92	0.00	-	-	**0.65**	**-0.81**	**0.25**	**0.041**	**0.051**	**0.65**	**-0.81**	**0.25**	**0.032**	**0.039**
10	1.00	-0.73	0.00	1.00	-0.83	0.00	-	-	1.00	-0.73	0.00	-	-	1.00	-1.38	0.00	-	-
11	1.00	-0.72	0.00	1.00	-0.72	0.00	-	-	**0.05**	**-0.61**	**0.25**	**0.039**	**0.089**	**0.05**	**-0.61**	**0.25**	**0.038**	**0.080**
12	1.00	-0.66	0.00	1.00	-0.64	0.00	-	-	1.00	-0.66	0.00	-	-	1.00	-1.32	0.00	-	-
13	1.00	-0.57	0.00	1.00	-0.62	0.00	-	-	1.00	-0.57	0.00	-	-	**0.12**	**-0.52**	**0.25**	**0.038**	**0.075**
14	1.00	-0.42	0.00	1.00	-0.61	0.00	-	-	1.00	-0.42	0.00	-	-	1.00	-1.15	0.00	-	-
15	1.00	-0.38	0.00	**0.30**	**0.15**	**0.25**	**0.042**	**0.080**	**0.30**	**0.15**	**0.25**	**0.041**	**0.100**	**0.30**	**0.15**	**0.25**	**0.037**	**0.061**
16	1.00	-0.37	0.00	1.00	-0.55	0.00	-	-	1.00	-0.37	0.00	-	-	1.00	-1.05	0.00	-	-
17	1.00	-0.28	0.00	1.00	-0.53	0.00	-	-	1.00	-0.28	0.00	-	-	**0.58**	**-0.26**	**0.25**	**0.039**	**0.052**
18	1.00	0.03	0.00	1.00	-0.46	0.00	-	-	1.00	0.03	0.00	-	-	1.00	-0.90	0.00	-	-
19	1.00	0.06	0.00	1.00	-0.33	0.00	-	-	**0.54**	**-0.07**	**0.25**	**0.048**	**0.084**	**0.54**	**-0.07**	**0.25**	**0.041**	**0.058**
20	1.00	0.10	0.00	1.00	-0.32	0.00	-	-	1.00	0.10	0.00	-	-	1.00	-0.71	0.00	-	-
21	1.00	0.10	0.00	1.00	-0.25	0.00	-	-	**1.00**	**-0.25**	**0.25**	**0.060**	**0.050**	**1.58**	**0.23**	**0.25**	**0.053**	**0.044**
22	1.00	0.11	0.00	1.00	-0.23	0.00	-	-	1.00	0.11	0.00	-	-	1.00	-0.40	0.00	-	-
23	1.00	0.15	0.00	1.00	-0.16	0.00	-	-	1.00	0.15	0.00	-	-	**1.61**	**0.50**	**0.25**	**0.059**	**0.046**
24	1.00	0.21	0.00	1.00	0.08	0.00	-	-	1.00	0.21	0.00	-	-	1.00	-0.15	0.00	-	-
25	1.00	0.25	0.00	**1.94**	**0.88**	**0.25**	**0.066**	**0.049**	**1.94**	**0.88**	**0.25**	**0.069**	**0.053**	**1.94**	**0.88**	**0.25**	**0.062**	**0.047**
26	1.00	0.49	0.00	1.00	0.25	0.00	-	-	1.00	0.49	0.00	-	-	1.00	0.14	0.00	-	-
27	1.00	0.55	0.00	1.00	0.26	0.00	-	-	1.00	0.55	0.00	-	-	**1.43**	**1.15**	**0.25**	**0.068**	**0.056**
28	1.00	0.67	0.00	1.00	0.56	0.00	-	-	1.00	0.67	0.00	-	-	1.00	0.42	0.00	-	-
29	1.00	0.71	0.00	1.00	0.75	0.00	-	-	**1.60**	**0.71**	**0.25**	**0.073**	**0.060**	**1.60**	**0.71**	**0.25**	**0.070**	**0.053**
30	1.00	0.76	0.00	1.00	0.79	0.00	-	-	1.00	0.76	0.00	-	-	1.00	0.46	0.00	-	-
31	1.00	1.13	0.00	1.00	0.83	0.00	-	-	**1.70**	**1.13**	**0.25**	**0.074**	**0.061**	**1.70**	**-**	**0.25**	**0.070**	**0.054**
32	1.00	1.17	0.00	1.00	1.05	0.00	-	-	1.00	1.17	0.00	-	-	1.00	1.14	0.00	-	-
33	1.00	1.47	0.00	1.00	1.12	0.00	-	-	1.00	1.47	0.00	-	-	**1.84**	**0.43**	**0.25**	**0.058**	**0.045**
34	1.00	1.53	0.00	1.00	1.18	0.00	-	-	1.00	1.53	0.00	-	-	1.00	1.24	0.00	-	-
35	1.00	1.63	0.00	**1.85**	**0.53**	**0.25**	**0.060**	**0.046**	**1.85**	**0.53**	**0.25**	**0.059**	**0.046**	**1.85**	**0.53**	**0.25**	**0.061**	**0.046**
36	1.00	1.68	0.00	1.00	1.41	0.00	-	-	1.00	1.68	0.00	-	-	1.00	1.39	0.00	-	-
37	1.00	1.71	0.00	1.00	1.47	0.00	-	-	1.00	1.71	0.00	-	-	**1.75**	**0.69**	**0.25**	**0.062**	**0.048**
38	1.00	1.76	0.00	1.00	1.58	0.00	-	-	1.00	1.76	0.00	-	-	1.00	1.57	0.00	-	-
39	1.00	1.93	0.00	1.00	1.68	0.00	-	-	**1.84**	**1.93**	**0.25**	**0.065**	**0.049**	**1.84**	**-**	**0.25**	**0.064**	**0.053**
40	1.00	1.98	0.00	1.00	1.77	0.00	-	-	1.00	1.98	0.00	-	-	1.00	1.99	0.00	-	-

Verilerin analizinde R 3.3.2'de yer alan "mirt" paketi kullanılmıştır. Öncelikle GM=CM ve GM>CM olduğu durumlar için her indeksin *p* değerleri hesaplanmıştır. Modeller karşılaştırılırken GM'nin parametre sayısı CM'nin parametre sayısına eşit veya daha yüksek olacak şekilde ayarlanmıştır. Maddelerin uyum gösterip göstermediği, 0.05 anlamlılık düzeyi kullanılarak belirlenmiştir. Eğer uyum indeksinin *p* değeri 0.05'den küçük ise madde modele uyumsuz olarak yorumlanmıştır. R 3.3.2 yazılımı kullanılarak her bir indeks için I. tip hata ve güç oranları hesaplanmıştır. İndekslerin I. tip hatalarının oranı, uyum göstermesi gerekirken uyumsuzluk gösteren madde sayısının toplam uyumlu olması gereken madde sayısına oranı ile hesaplanmıştır. İndekslerin güç oranları ise, uyumsuz olması gereken madde sayısı ile uyumsuz olan madde sayısının oranı ile hesaplanmıştır. GM=CM olduğu durumlarda I. tip hata ve GM>CM olduğu durumlarda güç oranları hesaplanmıştır. Bu simülasyon deseni Tablo 3'te verilmiştir. Her uyum indeksinin tüm koşullardaki I. tip hataları ve güç oranları hesaplanarak bu indeksler karşılaştırılmıştır. İndekslerin her faktör için temel ve ortak etkileri grafikleştirilmiştir.

Tablo 3. Çalışmada Kullanılan Simülasyon Deseni

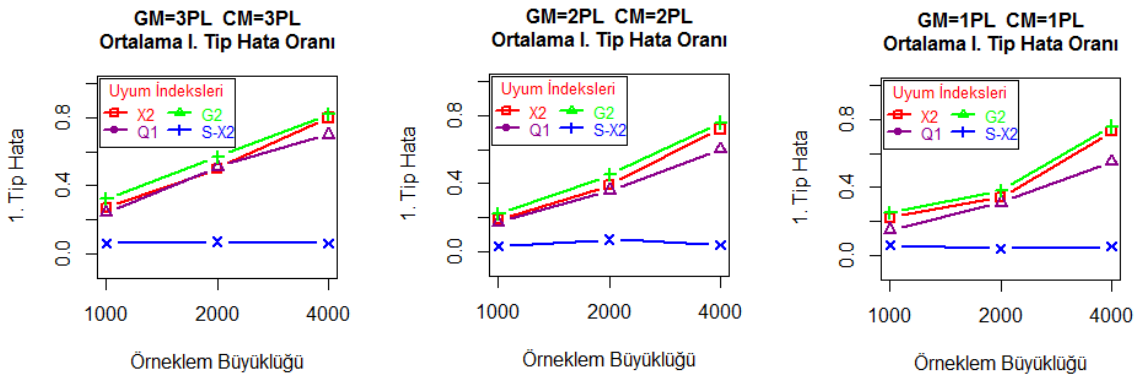
		ÜRETİLEN MODEL (GM)		
		1PL	2PL	3PL
ANALİZ MODEL (CM)	1PL	I. Tip Hata	Güç	Güç
	2PL	-	I. Tip Hata	Güç
	3PL	-	-	I. Tip Hata

BULGULAR

Aşağıdaki bölümde araştırma soruları çerçevesinde elde edilen bulgulara yer verilmiştir.

Örneklem Büyüklüğünün Madde-Uyum İndekslerinin I. Tip Hata ve Güç Oranlarına Temel Etkisine Ait Bulgular

Örneklem büyüklüğünün madde-uyum indekslerinin I. tip hata oranlarına ait temel etki grafikleri Şekil 1'de gösterilmiştir.



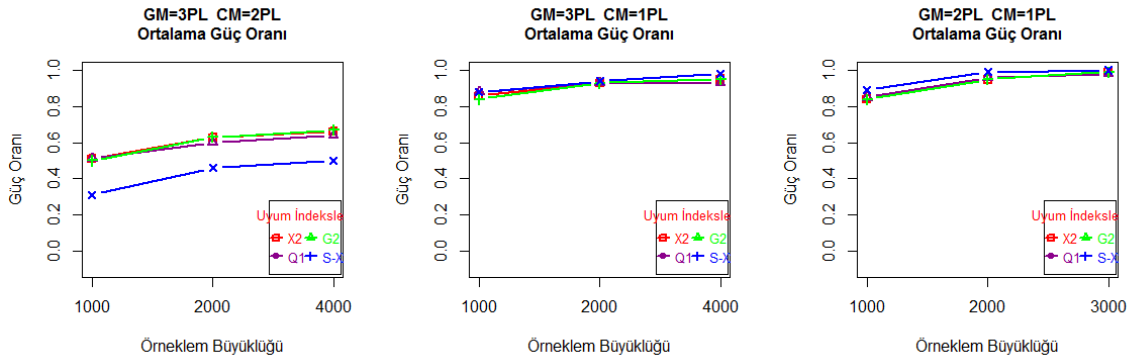
(a) (b) (c)

Şekil 1. Örneklem Büyüklüğünün Madde-Uyum İndekslerinin I. Tip Hata Oranlarına Temel Etkisi

Şekil 1 incelendiğinde örneklem büyüklüğü artıkça GM=3PL CM=3PL, GM=2PL CM=2PL ve GM=1PL CM=1PL durumları için $S\text{-}\chi^2$ indeksi hariç diğer indekslerin I. tip hata oranlarının arttığı

görülmektedir. χ^2, G^2 ve Q_1 indekslerinin I. tip hataları örneklem büyüklüğünün artması ile önemli ölçüde artış gösterirken, $S\text{-}\chi^2$ indeksinin I. tip hatasında önemsiz sayılabilecek değişimler olduğu görülmektedir. Tüm örneklem büyüklüklerinde en düşük I. tip hataya $S\text{-}\chi^2$ indeksinin sahip olduğu söylenebilir. $S\text{-}\chi^2$ indeksi ile geleneksel indeksler arasındaki I. tip hata oranları farkının 4000 örneklem büyüklüğünde önemli ölçüde artış gösterdiği görülmektedir.

Örneklem büyüklüğünün madde-uyum indekslerinin güç oranlarına etkisi ise Şekil 2'de gösterilmiştir.



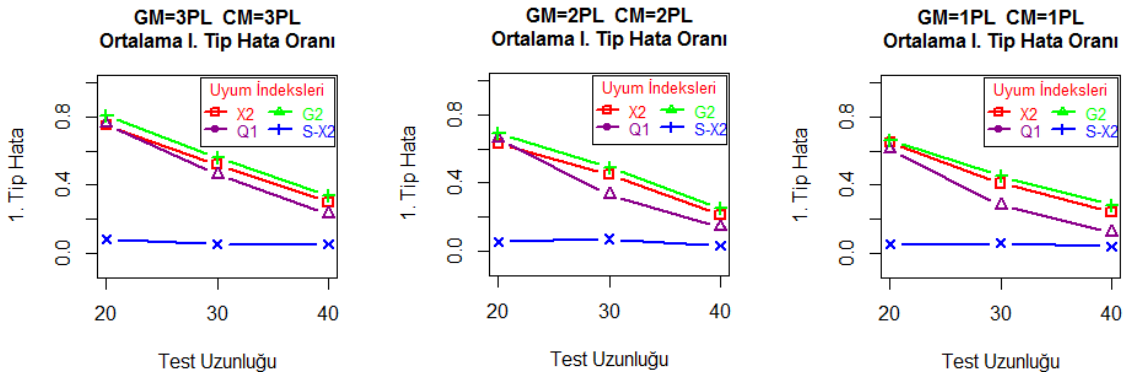
(a) (b) (c)

Şekil 2. Örneklem Büyüklüğünün Madde-Uyum Indekslerinin Güç Oranlarına Temel Etkisi

Şekil 2 incelendiğinde, örneklem büyüklüğü artıkça madde uyum indekslerinin güç oranları artmaktadır. Üretilen modelin, analiz edilen modele göre parametre farkı arttıkça indekslerin güç oranları artmakta ve birbirine yaklaşmaktadır. Şekil 2-a'da GM=3PL ve CM= 2PL durumu için en düşük güce $S\text{-}X^2$ indeksi sahipken, GM=3PL CM=1PL ve GM=2PL CM=1PL olduğu durumlarda tüm örneklem büyüklüklerinde indekslerin güç oranları arasında çok fazla fark bulunmamakla birlikte, $S\text{-}X^2$ indeksinin gücünün diğer indeksler gibi artış gösterdiği ve bu indeksin tüm örneklem büyüklüklerinde diğer indekslerin güç oranlarından daha yüksek olduğu görülmektedir. Şekil 2-a'ya göre GM=3PL ve CM=2PL olduğunda, indekslerin uyumsuz madde tespit etme güç oranlarının orta seviyede (0.30-0.60 arası), Şekil 2-b ve c'de ise, indekslerin uyumsuz maddeleri tespit etme güç oranlarının yüksek seviyede (0.80-1.00 arası) olduğu görülmektedir.

Test Uzunluğunun Madde-Uyum İndekslerinin I. Tip Hata ve Güç Oranlarına Temel Etkisine Ait Bulgular

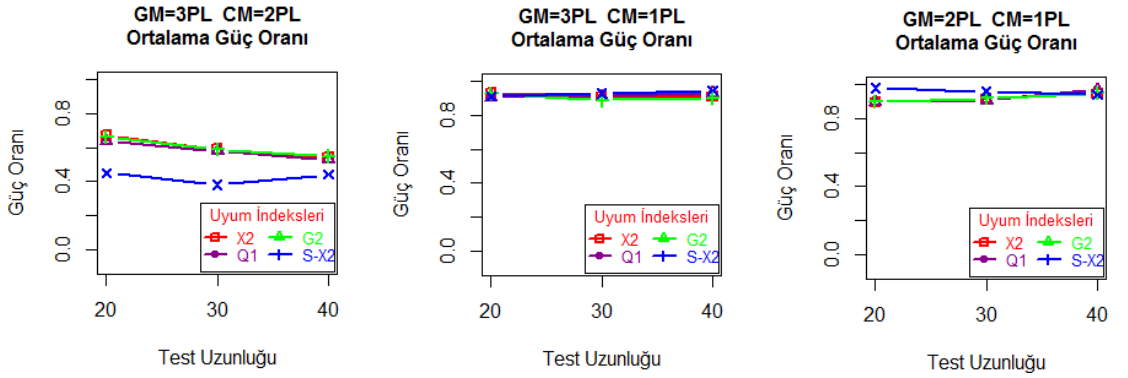
Test uzunluğunun madde-uyum indekslerinin I. tip hata oranlarına temel etkisi Şekil 3'te gösterilmiştir. Şekil 3 incelendiğinde tüm GM=CM durumları için test uzunluğu artıkça indekslerin I. tip hata oranlarının azaldığı görülmektedir. χ^2, G^2 ve Q_1 indekslerinin I. tip hata oranları, test uzunluğunun artmasıyla önemli ölçüde azalırken, $S\text{-}\chi^2$ indeksinin I. tip hata oranında çok az bir azalmanın olduğu görülmektedir. Tüm test uzunluklarında en düşük I. tip hata oranına ise $S\text{-}\chi^2$ indeksinin sahip olduğu söylenebilir. Kısa testlerde $S\text{-}\chi^2$ indeksi ile geleneksel indeksler arasındaki I. tip hata oranları farkı fazla iken, test uzunluğu arttıkça indekslerin I. tip hata oranları arasındaki farkın da azaldığı görülmektedir. Tüm GM=CM olduğu koşullarda test uzunluğu 40 iken indekslerin I. tip hata oranları yakındır.



(a) (b) (c)

Şekil 3. Test Uzunluğunun Madde-Uyum Indekslerinin I. Tip Hata Oranlarına Temel Etkisi

Test uzunluğunun madde-uyum indekslerinin güç oranlarına etkisi Şekil 4'te gösterilmiştir. Şekil 4-a incelendiğinde test uzunluğu artıkça GM=3PL CM=2PL durumu için S-χ^2 indeksi dışında madde-uyum indekslerinin güç oranları azalmaktadır. İndekslerin güç oranları 0.40 - 065 değerleri arasında değişmektedir. Ancak S-χ^2 indeksinin gücü 40 maddelik testte artış göstermiştir. Şekil 4-b ve 4-c incelendiğinde ise tüm indekslerin güç oranlarının tüm test uzunluklarında yüksek olduğu görülmektedir. İndekslerin güç oranları tüm maddelerde 0.90 - 0.95 aralığında değişmektedir. Tüm indekslerin gücü test uzunluğu arttıkça küçük değişimler göstermektedir.

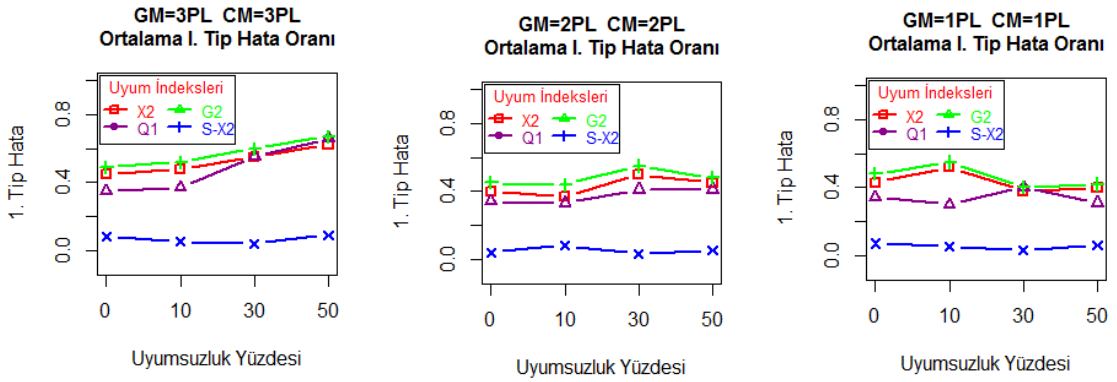


(a) (b) (c)

Şekil 4. Test Uzunluğunun Madde-Uyum İndekslerinin Güç Oranlarına Temel Etkisi

Uyumsuzluk Yüzdesinin Madde-Uyum İndekslerinin I. Tip Hata ve Güç Oranlarına Temel Etkisine Ait Bulgular

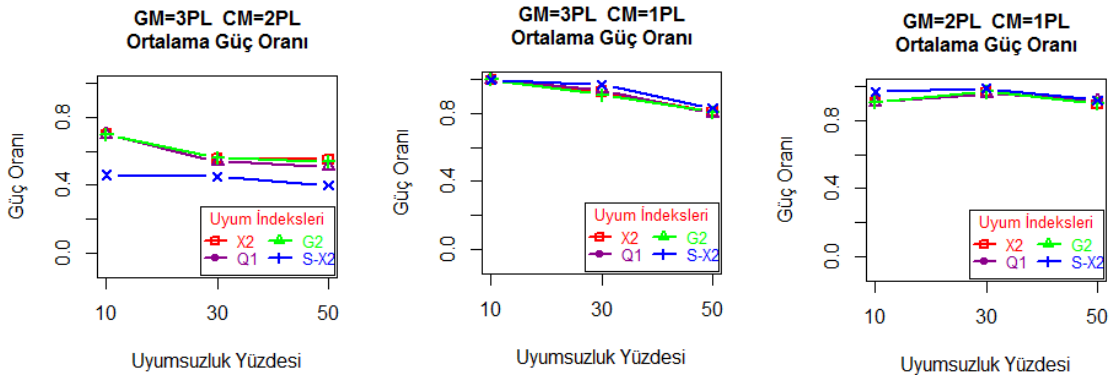
Uyumsuzluk yüzdesinin madde-uyum indekslerinin I. tip hatasına etkisi Şekil 5'te gösterilmiştir. Şekil 5-a incelendiğinde uyumsuzluk yüzdesi artıkça tüm durumlar için geleneksel indekslerin I. tip hatalarının kısmen arttığı, *S*-χ^2 indeksinin I. tip hata oranının ise küçük değişimler gösterdiği görülmektedir. Uyumsuzluk yüzdesi arttıkça I. tip hata oranı en fazla artan indeksin Q_1 indeksi olduğu görülmektedir. Ancak Şekil 5-b ve 5-c incelendiğinde, indekslerin I. tip hata oranlarının uyumsuzluk yüzdesinin artması ile düzensiz küçük değişimler gösterdiğini söyleyebiliriz. Tüm uyumsuzluk yüzdesinde *S*-χ^2 indeksi en düşük I. tip hata oranına sahiptir. Uyumsuzluk yüzdesi 50 olduğu durumlarda, *S*-χ^2 indeksi diğer indekslere göre önemli ölçüde, daha düşük I. tip hata oranına sahiptir.



(a) (b) (c)

Şekil 5. Uyumsuzluk Yüzdesinin Madde-Uyum İndekslerinin I. Tip Hatalarına Temel Etkisi

Uyumsuzluk yüzdesinin madde-uyum indekslerinin güç oranlarına etkisi Şekil 6'da gösterilmiştir. Şekil 6 incelendiğinde uyumsuzluk yüzdesi artıkça tüm durumlar için madde-uyum indekslerinin güç oranları küçük miktarda azalmaktadır. GM=3PL CM=2PL durumu için indekslerin güç oranları 0.40-065 değerleri arasında değişmektedir. Şekil 6-a'da geleneksel indekslerin güç oranları uyumsuzluk yüzdesi 10'dan 30'a artırıldığında hızlı düşüş gösterirken, 30-50 aralığında düşük miktarda azalmıştır. GM=3PL CM=2PL olduğu durumda $S\text{-}\chi^2$ tüm uyumsuzluk yüzdesinde en düşük güç oranına sahiptir. GM=3PL CM=1PL, GM=2PL CM=1PL durumunda tüm indekslerin güç oranları ise yüksek bulunmuş ve indekslerin güç oranları tüm maddelerde 0.80-1.00 aralığında değiştiği görülmüştür. Şekil 6-b ve 6-c incelendiğinde uyumsuzluk yüzdesi arttıkça indekslerin güç oranlarında az da olsa azalmanın olduğu görülmektedir.



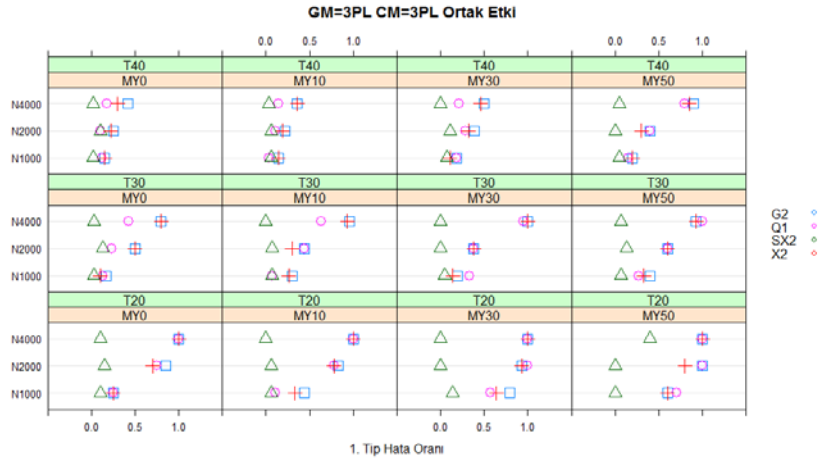
(a) (b) (c)

Şekil 6. Uyumsuzluk Yüzdesinin Madde-Uyum İndekslerinin Güç Oranlarına Temel Etki Grafiği

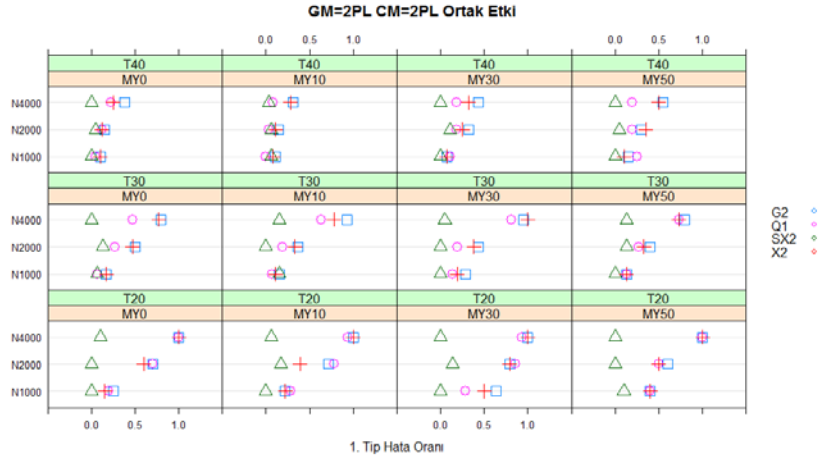
Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin I. Tip Hatalarına Ortak Etkisi

Şekil 7-a incelendiğinde (GM=CM=3PL), tüm koşullarda $S\text{-}\chi^2$ indeksinin I. tip hatası geleneksel indekslerin I. tip hatalarına göre daha düşüktür. Geleneksel indekslerin I. tip hataları örneklem büyüklüğü ve uyumsuzluk yüzdeleri arttıkça artmakta, test uzunluğu arttıkça ise I. tip hataları azalmaktadır. Geleneksel indeksler test uzunluğu 40 ve örneklem büyüklüğü 1000 olduğu

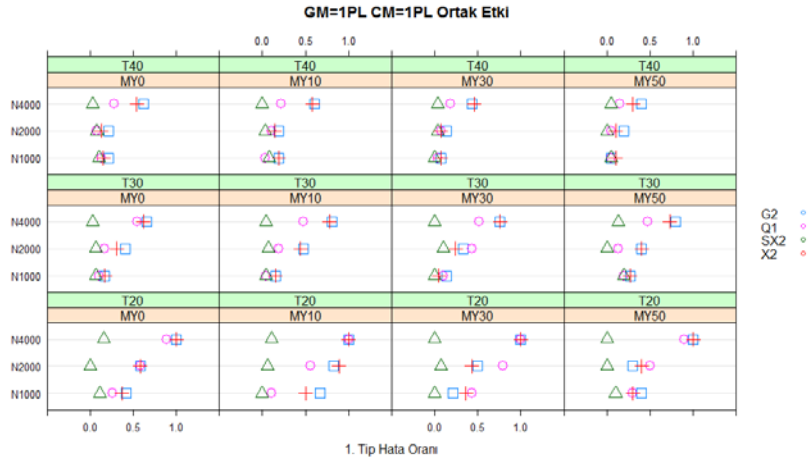
durumlarda $S\text{-}\chi^2$ indeksi ile benzer ve düşük I. tip hataya sahiptir. Ayrıca geleneksel indeksler uyumsuzluk yüzdesinin 0, örneklem büyüklüğünün 1000 olduğu tüm test uzunluklarında $S\text{-}\chi^2$



(a)

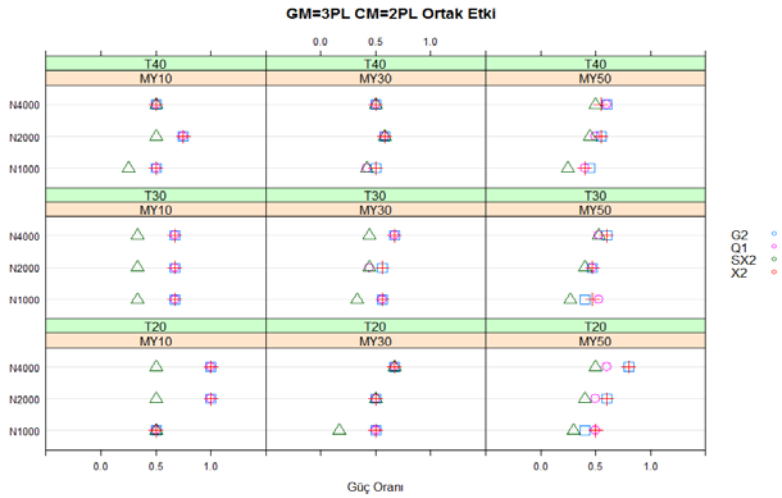


(b)

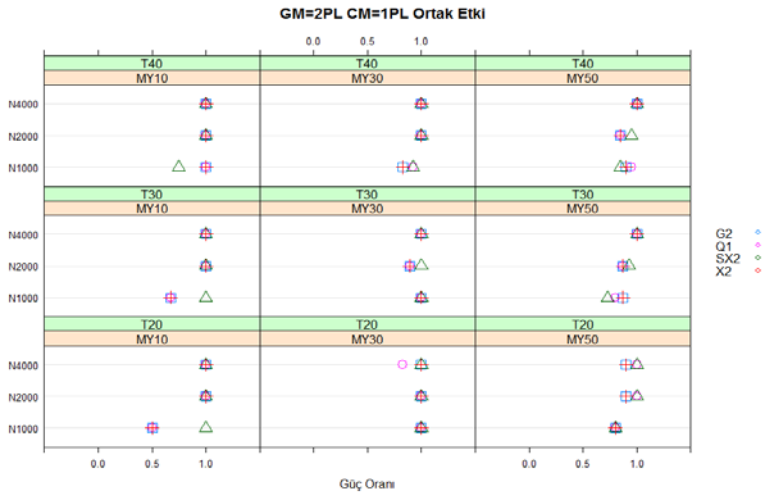


(c)

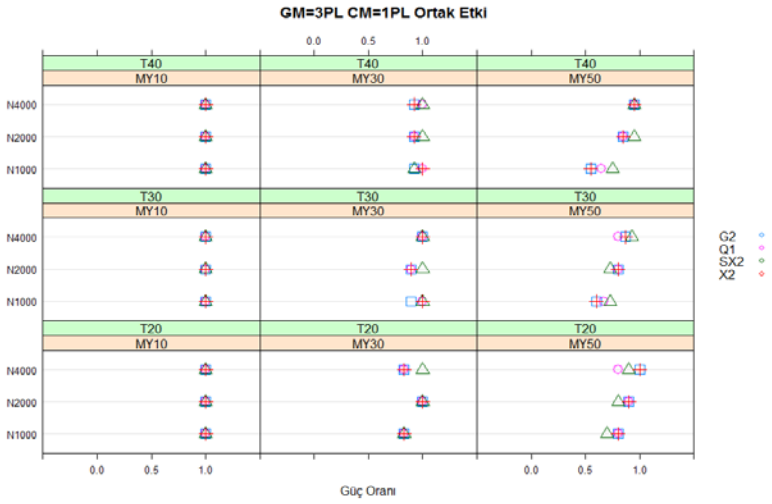
Şekil 7. Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin I. Tip Hatalarına Ortak Etkisi



(a)



(b)



(c)

Şekil 8. Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin Güç Oranlarına Ortak Etkisi

indeksine benzer ve düşük I. tip hata oranına sahiptir. Diğer tüm durumlarda $S\text{-}\chi^2$ indeksi daha düşük I. tip hata oranına sahiptir. Geleneksel indeksler, büyük örneklemlerde (2000 ve üzeri) ve yüksek uyumsuzluk yüzdesinde (%50) $S\text{-}\chi^2$ indeksine göre çok büyük I. tip hata oranına sahip olduğu görülmektedir. Benzer sonuçlar GM=CM=2PL ve GM=CM=1PL olduğu durumda da görülmektedir.

Örneklem Büyüklüğü, Test Uzunluğu ve Uyumsuzluk Yüzdesi Faktörlerinin Madde-Uyum İndekslerinin Güç Oranlarına Ortak Etkisi

Örneklem büyüklüğünün, test uzunluğunun ve uyumsuzluk yüzdesinin madde-uyum indekslerinin güç oranlarına ortak etkisi incelendiğinde, uyumsuzluk yüzdesinin düşük olduğu durumlarda, test uzunluğu ve örneklem büyüklüğünün tüm düzeylerinde, indekslerin güç oranlarında önemli değişiklikler olmadığı Şekil 8'de görülmektedir. Ancak uyumsuzluk yüzdesi arttıkça, yöntemlerin güç oranlarında azalmaların olduğu, özellikle düşük örneklemlerde ve uyumsuzluk yüzdesinin yüksek olduğu durumlarda, indekslerin uyumsuz maddeleri tespit etme güçlerinin azaldığı söylenebilir. Şekil 8 incelendiğinde, test uzunluğunun değişimlenmesi diğer faktörlerin tüm düzeylerinde, indekslerin güç oranlarında anlamlı değişimler yapmadığı söylenebilir. Şekil 8-b ve 8-c incelendiğinde, CM=1PL olduğu durumlarda, indekslerin tüm koşullarda yüksek ve benzer güce sahip oldukları görülmektedir.

SONUÇLAR ve TARTIŞMA

Bu çalışmada, Madde Tepki Kuramı'na göre, ikili puanlanan ve bir, iki ve üç parametreli lojistik modellere uygun olarak üretilen maddelerde, çeşitli madde-uyum indekslerinin çeşitli koşullardaki (örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdesi) I. tip hata ve güç oranlarının incelenerek, hangi koşulda hangi indeksin daha iyi sonuç verdiğini tespit etmek amaçlanmıştır. Bu amaçla simülasyon koşulları oluşturularak veriler üretilmiş ve sonuçları analiz edilmiştir.

Çalışmada yer alan koşullardan biri olan örneklem büyüklüğünün madde-uyum indekslerinin I. tip hata ve güç oranına etkisini incelemek için, 1000, 2000 ve 4000 olmak üzere üç farklı şekilde değişimlenmiştir. Madde-uyum indekslerinin I. tip hataları GM=CM olduğu durumlara göre değerlendirilmiştir. Orlando ve Thissen (2003), χ^2 ve Q_1 indekslerinin 500 ve 1000 örneklem büyüklüklerinde makul I. tip hataya sahip olduğunu ancak 2000 örneklem ve üzerinde I. tip

hatalarının yüksek olduğunu belirtmiştir. Ayrıca aynı çalışmada $S\text{-}\chi^2$ indeksinin 1000 örneklem büyüklüğü ve üzerinde makul hata değerine sahip olduğunu belirtmiştir. Ayrıca Orlando ve Thissen (2000), Stone ve Zhang (2003) ve Chon ve Dunbar (2010) yapmış oldukları çalışmada G^2 indeksinin 1000 ve üzeri örneklem büyüklüğünde, hata miktarının arttığını ve $S\text{-}\chi^2$ indeksinin daha makul I. tip hataya sahip olduğunu belirtmiştir. Bu çalışmada, GM=CM tüm durumları için, geleneksel indeksler olarak sınıflandırılan χ^2, Q_1 ve G^2 indekslerin I. tip hataları örneklem büyüklüğünden etkilenmiştir. Örneklem büyüklüğü arttıkça geleneksel indekslerin I. tip hata oranları da artmıştır. Ancak çalışmada alternatif indeks olarak yer alan $S\text{-}\chi^2$ indeksi örneklem büyüklüğünün değişimlenmesinden önemli ölçüde etkilenmemiştir. Tüm örneklem büyüklüğü düzeylerinde en düşük hataya $S\text{-}\chi^2$ indeksi sahiptir. Ayrıca geleneksel indeksler 1000 örneklem büyüklüğünün üzerinde yüksek hata oranına sahip olduğu görülmektedir. Elde edilen bu sonuçlar Orlando ve Thissen (2000), Orlando ve Thissen (2003), Stone ve Zhang (2003) ve Chon ve Dunbar'ın (2010) çalışmalarından elde edilen sonuçlarla benzerlik göstermektedir. Örneklem büyüklüğünün, madde-uyum indekslerinin güç oranlarına etkisi, GM=3PL CM=2Pl, GM=3PL CM=1PL ve GM=2PL CM=1PL durumları için incelenmiştir. Wells ve Bolt (2008), çalışmalarında örneklem büyüklüğünün artması ile madde-uyum indekslerinin gücünün arttığını göstermiştir. Bu çalışma sonucunda Wells ve Bolt'un (2008), çalışmasına paralel şekilde örneklem büyüklüğünün artması sonucunda indekslerin uyumsuz maddeleri tespit etme oranlarının arttığı soncuna ulaşılmıştır. Orlando ve Thissen (2000) çalışmasında, GM=3PL veya GM=2PL ve CM=1PL olduğu durumlarda indekslerin uyumsuz maddeleri tespit etme yüzdesinin %50 civarında olduğunu belirtmiştir. Ayrıca Orlando ve Thissen (2000) ve Stone ve Zhang (2003) GM=3PL ve CM=2PL olduğunda indekslerin uyumsuz maddeleri tespit etme oranlarının düştüğünü ifade etmiştir. Bu çalışmada GM=3PL ve CM=2PL olduğu durumlarda, güç oranlarının Orlando ve Thissen'in (2000) çalışmalarının sonucuna benzerlik gösterdiği söylenebilir. Ancak çalışma sonucunda GM=3PL veya GM=2PL ve CM=1PL olduğu durumlarda indekslerin güç oranlarının arttığı ve 0.80-1.00 aralığında değiştiği söylenebilir. Bu durumun, uyumsuz maddeler oluşturulurken parametrelerin uç değerlerde farklılaştırılmasından yani uyumsuzluk miktarının (misfit magnitude) değerinin yüksek olmasından kaynaklandığı düşünülmektedir.

Çalışmada yer alan bir diğer faktör olan test uzunluğunun madde-uyum indekslerinin I. tip hatalarına etkisi için madde sayısı 20, 30 ve 40 olmak üzere üç farklı düzeyde incelenmiştir. Çalışmada test uzunluğunun artması $S\text{-}\chi^2$ indeksinin I. tip hata oranını önemli ölçüde değiştirmez iken, geleneksel indekslerin I. tip hata oranını ise azalttığı sonucuna ulaşılmıştır. Mislev ve Bock (1990), G^2 indeksinin 20'den daha uzun testler için daha düşük I. tip hataya sahip olduğunu belirtmiştir. Ancak bu çalışma sonuçlarına göre, kısa testlerde geleneksel indeksler çok yüksek I. tip hata oranına sahip olduğu bulunmuştur. Ancak uzun testlerde (40 madde ve üzeri) geleneksel indeksler daha makul I. tip hata değerlerine sahip olduğu sonucuna ulaşılabilir. Bu çalışmada test uzunluğunun değişimlenen tüm düzeylerinde en küçük I. tip hata oranına $S\text{-}\chi^2$ indeksi sahiptir. Test uzunluğunun artırılması, madde-uyum indekslerin güç oranlarını önemli ölçüde etkilememektedir. GM=3PL CM=2PL olduğu durumlarda test uzunluğu arttıkça madde-uyum indeksinin gücü azalmaktadır. Ayrıca bu durum için indekslerin güç oranı orta düzeydedir. GM=3PL CM=2PL iken $S\text{-}\chi^2$ en düşük güce sahip indekstir. Ancak GM=3PL CM=1PL ve GM=2PL CM=1PL olduğu durumlar için indekslerin güç oranları test uzunluğunun artması ile önemli ölçüde değişmemektedir. Ayrıca indekslerin güç oranları test uzunluğunun tüm koşullarında yüksek bulunmuştur (0.90-1.00 arası). Elde edilen bu sonuçlar Orlando ve Thissen (2000, 2003), Stone ve Zhang'in (2003) çalışmalarının sonuçları ile tutarlılık göstermektedir.

Çalışma kapsamında çeşitli uyumsuzluk yüzde değerleri kullanılarak, uyumsuz madde yüzdesinin, madde-uyum indekslerinin, uyumsuz maddeleri tespit etme yeteneklerini nasıl etkilediği incelenmiştir. Uyumsuzluk yüzdesinin, indekslerin I. tip hatalarına etkisi incelendiğinde, uyumsuz madde sayısı arttıkça indekslerin I. tip hata oranları önemli bir şekilde değişmemektedir. Ancak GM=3PL CM=2PL olduğu durumda, yüksek uyumsuzluk yüzdesinde geleneksel indekslerin I. tip hata oranları artmaktadır soncuna ulaşılmıştır. Tüm uyumsuzluk yüzdelerinde, en düşük I. tip hataya $S\text{-}\chi^2$ indeksi sahiptir. Uyumsuzluk yüzdesinin, indekslerin güç oranlarına etkisi incelendiğinde, uyumsuzluk yüzdesi arttıkça indekslerin güç oranlarında küçük azalmalar olmaktadır. Ancak bu

değerler çok yüksek değildir. Wells ve Bolt (2008), yapmış oldukları çalışmada, uyumsuzluk yüzdesinin, indekslerin I. tip hatalarına ve güçlerine önemli bir etkisi olmadığını belirtmiştir. Bu çalışma sonucunda, Wells ve Bolt'un (2008) çalışmalarının sonucuna benzer olarak, uyumsuzluk yüzdesinin, indekslerin I. tip hata ve güç oranlarına önemli bir temel etkisinin olmadığı sonucuna ulaşılmıştır.

Çalışmada yer alan faktörlerin temel etkisinin yanı sıra ortak etkileri de I. tip hata ve güç oranlarına göre incelenmiştir. Literatür incelendiğinde (McKinley ve Mills, 1985; Yen, 1981) χ^2, Q_1 ve G^2 indekslerinin performanslarının yakın olduğu belirtilmiştir. Bu çalışmada da bu indeksler benzer performanslar göstermişlerdir. Ayrıca Orlando ve Thissen (2003) χ^2 ve Q_1 indekslerin 40 ve üzeri test uzunluğu ve 500 ile 1000 örneklem büyüklüğünde I. tip hataların makul düzeyde olduğunu belirtmiştir. Benzer bir şekilde bu çalışmada, örneklem büyüklüğü 1000, test uzunluğu 40 ve uyumsuzluk yüzdesi 0 olduğu durumlarda geleneksel indeksler S-χ^2 indeksi ile benzer ve düşük I. tip hata değerine sahiptir. Ancak büyük örneklemlerde, kısa testlerde ve yüksek uyumsuzluk yüzdelerinde S-χ^2 indeksi diğer indekslere göre daha düşük I. tip hataya sahiptir. Uyumsuzluk yüzdesinin indekslerin I. tip hata ve güç oranlarına önemli ölçüde temel etkisi olmadığı belirtilmiştir. Ancak uyumsuzluk yüzdesinin çalışmada yer alan diğer faktörlerle ortak etkisi incelendiğinde yüksek uyumsuzluk yüzdesinde (% 50) S-χ^2 indeksinin güç oranı diğer indekslere göre artmaktadır. Ortak etki açısından incelendiğinde test uzunluğunun değişimlenmesinin indekslerin güç oranına anlamlı etki etmediği sonucuna ulaşılmıştır.

Sonuç olarak, geleneksel madde-uyum indeksleri küçük örneklemlerde (1000), uzun testlerde (40 madde) ve uyumsuzluk yüzdesinin çok yüksek olmadığı durumlarda yeterli sonuçlar vermektedir. Chone, Lee ve Ansley (2007) S-χ^2 indeksinin uyum iyiliği istatistikleri arasında madde uyumunu değerlendirmede iyi bir alternatif olduğunu belirtmiştir. Ayrıca Orlando ve Thissen (2000) 1000 ve Stone ve Zhang (2003) 2000 ve üzeri örneklem büyüklüğünde S-χ^2 indeksinin tercih edilebileceğini belirtmiştir. İlgili literatürdeki çalışmalara benzer şekilde, bu çalışma sonucunda, 1000 örneklem ve üzerinde, 20 ve üzeri test uzunluklarında S-χ^2 indeksi, diğer alternatif indekslere göre daha doğru sonuçlar vermektedir. Ayrıca S-χ^2 indeksi, uyumsuzluk yüzdesinin yüksek olduğu durumlarda madde-uyumunun değerlendirildiği çalışmalarda tercih edilebilir.

Bu çalışmada, örneklem büyüklüğü, test uzunluğu ve uyumsuzluk yüzdelerinin, χ^2, Q_1, G^2 ve S-χ^2 indekslerinin I. tip hata ve güç oranlarına etkisi incelenmiştir. Madde uyumunu etkileyen bu faktörlerin yanı sıra indekslerin I. tip hata ve güç oranlarına etki eden uyumsuzluk büyüklüğü ve uyumsuzluk yeri (misfit location) gibi faktörler çalışmaya dahil edilerek indekslerin I. tip hata ve güç oranlarına etkileri incelenebilir. Ayrıca Stone (2000) tarafından ölçeklenmiş düzenlenmiş χ^2* indeksi veya bayese dayalı madde-uyum indeksleri kullanılarak bu faktörler altında uyumsuz maddeleri tespit etme oranları hesaplanabilir. Bu çalışmada, ikili puanlanan maddeler için simülatif veriler kullanılmıştır. Çoklu puanlanan maddeler için hem gerçek hem de simülatif veriler kullanılarak farklı çalışmalar gerçekleştirilebilir. Bununla birlikte, kayıp verili ve MTK'nın sayıltılarının yerine getirilmediği durumlara, gerçek hayattan elde edilen verilerde sıklıkla karşılaşılmaktadır. Gerçek durumlara yakın veriler elde edilerek benzer indekslerin I. tip hata ve güç oranları incelenebilir.

KAYNAKÇA

Ames, A. J. (2015). *Bayesian model criticism: Prior sensitivity of the posterior predictive checks method* (Doctoral dissertation). University of North Carolina.

Ames, A. J., & Penfield, D. R. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*(3), 39–48.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Chon, K. H., Lee, W. C., & Ansley, T. N. (2007). *Assessing IRT model-data fit for mixed format tests*. Center for Advanced Studies in Measurement and Assessment CASMA Research Report, No: 26.

Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement, 47*(3), 318–338.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement, 65*, 42–50.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19,* 143-165.

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

Glas, C. A. W., & Su´arez Falc´on, J. C. (2003). A comparison of item-fit statistics for the three parameter logistic model. *Applied Psychological Measurement, 27*, 87–106.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* New York: Springer Science+Business Media, LLC.

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

Kang, T., & Chen, T. T. (2008). Performance of the generalized *S-*χ^2 item fit index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391-406.

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-χ^2 item fit index for the graded response model. *Asia Pacific Educ. Rev., 12,* 89–96.

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of Item Response Theory item fit indices for the graded response model. *Organizational Research Methods 14(1),* 10-23.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement,* 9, 49-57.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG-W. Item analysis and test scoring with binary logistic models*. Moresville, IN: Scientific Software.

Orlando, M. (1997). *Item fit in the context of Item Response Theory.* (Doctoral dissertation, University of North Carolina, 1997). Dissertation Abstracts International, 58/04-B, 2175.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous Item Response Theory models. *Applied Psychological Measurement, 24(1),* 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of *S-*χ^2: An item fit index for use with dichotomous Item Response Theory models. *Applied Psychological Measurement, 27*, 289-298.

Reise, S. P. (1990). A comparison of item-and person-fit methods of assessing model-data fit in IRT. *Applied Pyschological Measurement, 14(2),* 127-137.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*, 58-75.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of Item Response Theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40,* 331–352.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement, 35(4),* 280–295.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item Response Theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

von Schrader, S., Ansley, T. N., & Kim, S. (2004). *Examination of item fit indices for polytomous item response models.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in Item Response Theory. *Applied Measurement in Education, 21*(1), 22-40.

Wells, C. S., & Hambleton, R. K. (2016). *Model fit with residual analyses.* In W. J. van der Linden (Ed.) *Handbook of item response theory*. Volume two, Statistical tools (pp. 395-413). NewYork: CRC Press. Taylor ve Francis Group.

Wright, B., & Panchapakesan, N. A. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29*, 23–48.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262.

EXTENDED ABSTRACT

Introduction

Item Response Theory (IRT) is a strong scaling technique that uses mathematical models, which estimate performance of examinee, by using item and person's characteristics (Embretson & Reise,

2000). IRT has many advantages in interpreting and reporting the test scores. However these advantages are valid only test data are fitted to selected model (Hambleton & Swaminathan, 1985). Beside this, the property of parameter invariance of IRT models is one of the most important issues in IRT. Parameter invariance property of IRT models is achieved only when the model and data are fit. Thus model data fit plays important role for valid applications of IRT models to obtain decisive score scales. There are many item-fit indices to examine model data fit. These indices can be classified as traditional indices and alternative indices. Traditional indices such as χ^2, Q_1, G^2 examine model data fit by comparing observed performance with estimated performance in various ability groups for each item under chosen IRT model. In traditional indices, ability interval is usually determined arbitrarily. These arbitrary ability intervals can influence the results of the study (Orlando & Thissen, 2000; Reise, 1990). Observed responses are model-dependent since discrete intervals are depended to ability. On the contrast, in alternative indices such as $S\text{-}\chi^2$, $S\text{-}G^2$, individuals do not have to be grouped arbitrary interval along the ability. These indices are conditioned on total score instead of ability. In this study, χ^2, Q_1 and G^2 indices as traditional indices and $S\text{-}\chi^2$ as an alternative index were examined.

Method

The purpose of this study was to investigate Type I error and power rates of the item fit indices through various conditions (sample sizes, different test lengths and different magnitudes of misfit) for dichotomously generated items based on one-, two-, and three-parameter logistic models in Item Response Theory. Hence, it is expected to contribute theoretical studies related to item fit in dichotomous IRT models. Thus, this study is theoretical research in this respect. In this study, the type I error and power rates of these item fit indices were assessed in a simulation study. χ^2, Q_1 and G^2 indices as traditional item fit indices and $S\text{-}\chi^2$ index as alternative indices were assessed. The performance of four different item fit indices in study were compared by manipulating three different sample size (1000, 2000, 4000), three different test lengths (20, 30, 40) and four different misfit percentage (%0, %10, %30 and %50). Item responses were generated using the R 3.3.2 software program. Data were generated for both generating models (GM) and analysis models (CM) with respect to manipulated factors and level of these factors. For all GM, item discrimination parameters were are 1, b parameters were obtained from uniform distribution which minimum was -2 and maximum was 2 and c parameters were 0. For all CM, b parameters were generated by adding +-0.75 to b parameters in GM. For 2PL and 3 PL models in CM, item discrimination parameters were obtained from uniform distribution which minimum was 0 and maximum was 2. For 3PL in analysis model, c parameter were 0.25. Ability distribution of examinees was obtained from normal distribution which mean was 0 and standard deviation was 1. In three conditions, CM has fewer parameters than GM and in other three conditions; CM and GM have identical parameters. After data generation, item fit indices were analyzed by using "mirt" package in R software. The p value of item fit indices and their degrees of freedom were calculated for both item responses for GM and the item responses for CM. Type I errors and power rates of item fit indices were examined according to significance levels of 0.05. All item fit indices in this study were compared by calculating the type I error for conditions GM=CM and power rates for conditions GM>CM of each item fit indices under all conditions.

Results and Discussion

Sample size is an important factor to assess the item fit indices. The results of this study indicated that the performance of all item fit indices were influenced from sample sizes, however, the degree of effect of sample size differs in item fit indices. Type I errors of traditional indices inflated by increasing sample size. It is not suggested to use traditional indices when the sample size was larger than 1000. On the other hand, $S\text{-}\chi^2$ had acceptable Type I error across all sample size. These results are consistent with the results of the studies of Orlando & Thissen (2000), Orlando & Thissen (2003), Stone & Zhang (2003) and Chon & Dunbar (2010). Sample size has also effects on the

power rates of the item fit indices. Item indices performed well for detecting misfit items by increasing sample size. Test length is the one of the factors, which affects the Type I errors, and power rates of item fit indices. Type I errors of traditional item fit indices decreased by increasing test length, whereas, Type I errors of $S\text{-}\chi^2$ remained same. However, $S\text{-}\chi^2$ was more acceptable than traditional indices across all test lengths. Traditional indices and $S\text{-}\chi^2$ performed closely when the test length was 40. $S\text{-}\chi^2$ performed better than other indices for 20 and 30 test length. These results are identical to findings of studies of Orlando & Thissen (2000, 2003) and Stone & Zhang (2003). Percentage of misfit had no substantial influence on either Type I errors or power rates. These results are parallel with the results of the Wells and Bolt (2008) study. Overall, among the item fit indices, $S\text{-}\chi^2$ produced more accurate results across all conditions so that it is recommended for assessing item fit. Manipulating the test length did not affect the power rates of item fit indices in a way.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);98-111

A Comparison of IRT Vertical Scaling Methods in Determining the Increase in Science Achievement*+

Fen Başarısındaki Artışın Belirlenmesinde Madde Tepki Kuramına Dayalı Dikey Ölçekleme Yöntemlerinin Karşılaştırılması

Aylin ALBAYRAK SARI** Hülya KELECIOGLU ***

Abstract
This study is based on a vertical scaling implemented with reference to the Item Response Theory, and involves a comparison of vertical scaling results obtained through the application of proficiency estimation methods and calibration methods. The vertical scales thus developed were assessed with reference to the criteria of grade-to-grade growth, grade-to-grade variability, and the separation of grade distributions. The data used in the study pertains to a dataset composed of a total of 1500 students from twelve primary schools in the province of Ankara, characterized by different levels of socio-economic cultural development. The comparison of the findings pertaining to the first and the second sub-problems reveals that the mean differences found through separate calibration were lower than those applicable to concurrent calibration, while the standard deviation found in the case of separate calibration were again lower than the values established through concurrent calibration. Furthermore, the scale of impact in the case of separate calibration was again lower than the values applicable to concurrent calibration. The results reached for all three criteria, using the concurrent calibration method were ranked in the order ML < MAP < EAP, with ML leading to the lowest value while EAP producing the highest one. In case of separate calibration, on the other hand, the ranking of results was found to vary with reference to the criteria applied.

Key words: Item response theory, vertical scaling, calibration methods, proficiency estimation methods.

Öz
Bu araştırmada Madde Tepki Kuramına dayalı dikey ölçekleme çalışması yürütülmüş, kalibrasyon yöntemleri ve yetenek kestirim yöntemleri kullanarak elde edilen dikey ölçekleme sonuçları karşılaştırılmıştır. Elde edilen dikey ölçekler, bir sınıf düzeyinden diğer sınıf düzeyine olan büyüme, sınıf düzeyleri arasındaki çeşitlilik ve düzey dağılımlarının ayrımı kriterlerine göre değerlendirilmiştir. Çalışmanın verileri Ankara ili farklı sosyoekonomik kültüre sahip on iki ilköğretim okulundan toplam 1500 öğrenciden toplanmıştır. Birinci ve ikinci alt probleme ait elde edilen bulgular karşılaştırıldığında, ayrı kalibrasyon ile elde edilen ortalama farkların eş zamanlı kalibrasyon ile elde edilen ortalama farklarından daha düşük olduğu, ayrı kalibrasyon ile elde edilen standart sapma değerlerinin genel olarak eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu ve ayrı kalibrasyon ile elde edilen etki büyüklüğü değerlerinin eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu görülmektedir. Eş zamanlı kalibrasyon yöntemi ile her üç kriter için de elde edilen sonuçların ML < MAP < EAP şeklinde sıralandığı; en küçük değerlerin ML, en büyük değerlerin ise EAP ile elde edildiği görülmektedir. Ayrı kalibrasyon da ise sonuçların sıralamalarının kullanılan kriterlere göre farklılaştığı görülmektedir.

Anahtar Kelimeler: Madde tepki kuramı, dikey ölçekleme, kalibrasyon yöntemleri, yetenek kestirim yöntemleri.

* This study is a part of Aylin Albayrak Sarı's doctoral dissertation titled "A Comparison of IRT Vertical Scaling Methods in Determining of the Increase in Achievement of Science Education" and conducted under the supervision of Professor Hülya Kelecioğlu.

+ This study was supported by Hacettepe University Scientific Research Projects Coordination Unit (Project Nu: 014 T03 700 001-587).

** Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: aylinalb@hacettepe.edu.tr
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: hulyaebb@hacettepe.edu.tr

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 8, Sayı 1, İlkbahar 2017, 98-111.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 8, Issue 1, Spring 2017, 98-111.
DOI: 10.21031/epod.286221
Received: 17.01.2017
Accepted:03.03.2017

INTRODUCTION

Exams applied at schools serve for a wide range of objectives. When deciding on the school a student will attend, or setting the test score a candidate is expected to have for admission for a university, deciding on what to do to enhance the education system, and assessing the changes in educational practices, information derived from exams is used (Kolen, & Brennan, 2004).

In order to ascertain the level of change in academic development from one year to the next, developmental scale scores established by converting the scores pertaining to students at different levels of class into a common scale is used (Kolen, & Brennan, 2004). An awareness of the level of development through the years can provide dependable knowledge about the continuity of success, whereupon improvements at the student and class level can be effected. Large-scale assessments covering the period from K-12 grade involved numerous studies to assess the academic achievement levels of the students. It is necessary to develop a single scale score for all students' performances in all levels for reviewing and comparing academic development through the years and presenting all test scores in a single scale regardless of the year.

The fundamental problem regarding the level of academic development from one year to the next is the differences in the level of difficulty of tests, as well as their contents, even if the general topic may be the same. In order to overcome this issue, a common set of items are directed to students from consecutive years of education and the scores of students at different proficiency levels are converted into a common scale by using these items.

The process of establishing a link between the scores received in tests applied to different years is called vertical scaling (Kolen, & Brennan, 2004; McBridge, & Wise, 2001). The primary reason of applying scaling on test batteries is to provide a developmental scale score to the test developers to enable monitoring the progress in students' achievement levels (Loyd, & Hoover, 1980).

Different data collection designs, scaling methods, calibration methods, proficiency estimation methods or evaluation criteria can be applied in vertical scaling processes. The researchers would be required to make certain decisions about the designs and methods to be used in the scaling process. Such decisions were observed to have an impact on vertical scaling, and therefore the patterns indicating the change in the achievement levels of students (Tong, & Kolen, 2007). There is a brief discussion of the designs and methods chosen for this study.

Data Collection Designs

In equating, the data collection design is often called the "scaling design" (von Davier, & Wilson, 2008). Non-equivalent groups anchor test design, scaling design, and equal-to-group design are the most common used designs in vertical scaling. As the non-equivalent groups anchor test design is used in the present study, the following section will provide a brief description of the method.

The non-equivalent groups anchor test design enables the comparison of the performance of groups with reference to anchor items by building on the overlapping structure of test batteries in elementary education. For each grade, a test compatible with the level of the grade would be developed, and each such test would be applied only to the relevant grade. The test-takers' level of success with the anchor items are then used to establish the level of growth from one year to the next (Kolen & Brennan, 2004). As the design is applied on two non-equivalent groups, it is called non-equivalent groups anchor test (or anchor item) design (NEAT) (von Davier, Holland, & Thayer, 2004). Where anchor items are chosen correctly, this design helps reduce the equating error in the scaling (Hambleton, Swaminathan, & Rogers, 1991; Holland, & Dorans, 2006).

Scaling Design

Each equating method is based on a distinct theory and assumption. The equating methods are categorized as methods based on the Classical Test Theory (CTT) or on the Item Response Theory (IRT), with reference to the underlying theoretical framework.

Equating based on IRT involves the development of a mathematical relationship between the scores in two distinct forms of a test (Dongyang, 2009). Equating methods based on IRT are developed on the basis of the assumption of the existence of a mathematical function defining the relationship between the respondents' proficiency level (θ) and the probability to provide a correct response (Kolen & Brennan, 2004). Understanding, implementing, and explaining IRT methods are harder compared to CTT methods; yet IRT methods are more flexible (Harris, 2003).

One-parameter logistic model, two-parameter logistic model, and three-parameter logistic models may be applied with reference to the scale, in case of items scored on a binary scale (1-0). The present study applies a two-parameter logistic model (2-PLM).

Calibration Methods

When NEAT design is used in vertical scaling, the anchor items enable the establishment of a shared scale linking the test levels of different grades. With NEAT design, IRT parameters are either estimated for each test level by running the program separately, or estimated concurrently as the program is ran only once (Kolen, & Brenan, 2004). These calibration methods are called concurrent and separate calibration methods (Meng, 2007).

Concurrent calibration: Data pertaining to all grades is calibrated at once, to produce a vertical scale in concurrent calibration. The item parameters of the forms are estimated on the basis of the assumption that anchor items present the same item parameters for consecutive grades (Meng, 2007). In this context, the first thing to do is to set a reference grade, followed by the development of a scale with a mean of 0 and standard deviation of 1, pertaining to the scaled proficiency estimations for consecutive grades (Çetin, 2009). The item parameters for the anchor items included in the target test are estimated once again after adjustment to the values of the reference test. The item parameters pertaining to anchor items are known, while IRT calibrations are used to place non-anchor items of the target test with reference to the reference test scale (Meng, 2007).

Separate calibration: In separate calibration, the item parameters are calculated separately for each grade. As the item and proficiency parameters established separately for two different test forms have different scales, they are not readily comparable. With a view to enabling comparisons, a grade is chosen as the reference level, and θ scale is set as the starting scale for a grade. Item and proficiency parameters' estimation are used to place on the starting scale by using a series of linear conversions, with reference to the anchor items in the NEAT design (Kolen, & Brennan, 2004). Numerous linking procedures were developed in order to place the results obtained through the separate calibration on a single shared scale. The studies comparing various equating methods proposed in the literature recommend the use of Haebara and Stocking Lord (SL) methods utilizing item and test characteristics curves, instead of moment methods applying item parameters (Hanson, & Béguin, 2002; Kim, & Kolen, 2006; Kolen, & Brennan, 2004). Furthermore studies note that SL method generates less error compared to alternative methods (Hanson, & Béguin, 2002; Karkee, & Wright, 2004; Kim, 2007). Therefore, the present study applied Stocking Lord method as a characteristic curve equating method.

Furthermore, the present study compares the results obtained through scaling via both concurrent and separate calibration.

Proficiency Estimation Methods

Once the item parameters are converted into a common scale using an appropriate calibration method, the methods for estimating proficiency level should be decided. Total score or pattern scoring can be used when applying θ proficiency level estimation with reference to item response theory. The total score method, which offers a more practical and simpler approach, is used more frequently compared to the pattern scoring method. However, its error rate is larger compared to pattern scoring, while the amount of information it provides is smaller (Tong, & Kolen, 2010). For proficiency estimation regarding the binary items coded as 1-0 in IRT, often three distinct proficiency estimation methods are used. These are Maximum Likelihood (ML), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP) estimation methods. The present study provides a comparison of the results achieved through all three proficiency estimation methods.

Evaluation Criteria

The final stage of the scaling study involves the comparison of the results obtained. The normative characteristics of developmental scale scores constitute the subject matter of numerous studies. The characteristics of the scale scores are compared in order to be able to compare the results of the vertical scaling analysis. These characteristics refer to grade-to-grade growth, grade-to-grade variability, and separation of grade distributions. Grade-to-grade growth is assessed with reference to mean difference between consecutive grades, grade-to-grade variability is assessed with reference to standard deviation between consecutive grades, and separation of grade distributions are interpreted with reference to the effect size index proposed by Yen (1984) (Kim, 2007; Kolen, & Brennan, 2004). The present study provides a comparison of the results through all three evaluation criteria.

Purpose of the Study

The literature has not yet to come up with a common view about which method reveal the best and most accurate depiction of the increase in the level of the students' achievement. Nevertheless, vertical scaling is used by numerous test developers, and every test developer determine its own vertical scaling processes (Tong, & Kolen, 2007).

Vertical scaling as a means of revealing the development of students' achievement from one grade to the next, has subsequently became an important field, and there is an increase in the number of the vertical scaling studies. The present study can provide a model about monitoring of the development in terms of students' achievement levels.

A glance at the literature reveals the rarity of studies based on real data, while studies based on simulated data are more common. The present study, on the other hand, is based on the results of science achievement tests applied with 1500 students enrolled in six different schools. In this vein, the study is expected to contribute to the literature as a model based on real data.

The purpose of the study is to implement a vertical scaling analysis based on the item response theory, and to come up with a comparison of the developmental scale scores established through the application of calibration methods (separate and concurrent calibration) and estimation methods (maximum likelihood, maximum a posteriori, and expected a posteriori estimation), with reference to the mean, standard deviation and effect size. That is why the study discusses the grade-to-grade growth, grade-to-grade variability, and separation of grade distribution characteristics pertaining to developmental scale scores. Mean and mean differences were employed to assess grade-to-grade growth, standard deviation figures for each grade were used to assess the grade-to-grade variability, and effect size were analyzed to assess the separation of grade distribution.

Research Questions

This study maintains vertical scales over three forms and investigated the question "How does the evaluation criteria vary by using various calibration methods and proficiency estimation methods in terms of vertical scaling on the basis of item response theory?". Specifically, the research questions to be investigated in line with this problem statement are as below:

1. How do;
    a. grade-to-grade growth,
    b. grade-to-grade variability, and
    c. separation of grade distribution

vary with respect to maximum likelihood, maximum a posteriori, and expected a posteriori estimations using concurrent calibration?

2. How do;
    a. grade-to-grade growth,
    b. grade-to-grade variability, and
    c. separation of grade distribution

vary with respect to maximum likelihood, maximum a posteriori, and expected a posteriori estimations using separate calibration?

METHOD

Type of Study

Because the existing methods and techniques in the research were tested through real data, and since the aim was to contribute to theoretical studies by designating the methods with minimum error, the research is a fundamental study (Creswell, 2013).

Participants

The participants of the study consist of 6^{th}, 7^{th}, and 8^{th} grades. The data used in the study were gathered from a total of 1500 students from 12 distinct schools; two from each of the Altindag, Cankaya, Golbasi, Kecioren, Sincan, and Mamak districts of Ankara province.

The science achievement test applied was developed using items selected out of Placement Exam (SBS), High School Entrance Examination (OKS), and Free Boarding and Scholarship Examination (PYBS) applied between the years 2008-2012 by checking the item discrimination and item difficulty indices, whereupon the items were compiled to achievement tests of 40 items for each of the three grades. Ten items were identified as anchor items to enable chain scaling between consecutive grades. While Hambleton, Swaminathan and Rogers (1991) note that 20% of the overall test would be a sufficient guideline to establish the number of anchor items, many studies note that increase in the number of anchor items would help reduce the standard deviation regarding the assessment sought through the test (Boughton, Lorie, & Yao, 2005; Kim, Lee, Kim, & Kelley, 2009). Therefore, the present study employed an anchor item ratio of 25% of the total number of items.

Research Design

In this research, the non-equivalent groups anchor item design was used. Even though the design is one of the most frequently employed ones, it is also one of the most flexible and most complex

designs (Sinharay, & Holland, 2007). Even though it is a design preferred on practical grounds, it is also less restrictive compared to other designs (Zhu, 1998).

Data Analysis

Before running the analyses, data was subjected to preprocessing to remove incomplete or missing data from the dataset. Furthermore, the scores received from the science achievement test were checked for unidimensionality, local independence, and model-data fit compliance among major IRT assumptions.

When unidimensional Item Response Theory (IRT) is used for equating, it is necessary to test the unidimensionality assumption for the tests (Hambleton, & Swaminathan, 1985). In order to test the unidimensionality assumption of the item response theory, confirmatory factor analysis (CFA) was applied to all three grade levels of the science tests given to students, leading to the testing of the model for a significance level of 0.05. Numerous goodness of fit indices are used in order to evaluate the model-data fit. Among these, the most frequently used indices, namely Chi-Squared Test ($\chi 2$ / sd), Root Mean Square Error of Approximation (RMSEA), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), Comparative Fit Index (CFI), and Normed Fit Index (NFI) were checked. The obtained results are presented in Table 1.

Table 1. Good Fit Indices Calculated Through Confirmatory Factor Analyses for Science Test

Level of Fit	Perfect Fit Value	Allowable Fit Value	Model Value		
			6th Grade	7th Grade	8th Grade
χ^2 / sd	$0 < \chi^2 / sd \leq 2$	$2 < \chi^2 / sd \leq 5$	1.76	2.35	1.98
RMSEA	$0 < RMSEA < 0.05$	$0.05 < RMSEA < 0.10$	0.05	0.08	0.05
GFI	$0.95 \leq GFI \leq 1$	$0.90 \leq GFI \leq 0.95$	0.93	0.93	0.94
AGFI	$0.90 \leq AGFI \leq 1$	$0.85 \leq AGFI \leq 0.90$	0.92	0.90	0.95
CFI	$0.97 \leq CFI \leq 1$	$0.95 \leq CFI \leq 0.97$	0.97	0.97	0.98
NFI	$0.95 \leq NFI \leq 1$	$0.90 \leq NFI \leq 0.95$	0.97	0.94	0.95

(Ref.: Schermelleh-Engel, Moosbrugger & Müller, 2003)

A review of the goodness of fit indices obtained through CFA analysis and presented in Table 1 reveals that the model presents a high level of fit for all three grades, and the model meets the requirements of the unidimensionality assumption. Based on the CFA analysis, it can be said that data meets the unidimensionality assumption; hence the science achievement test assesses a single feature in all grades involved.

Local independence means that a response given to each item is independent from others, and the possibility of giving a positive answer to an item is not affected by other items. When the proficiency level is fixed, the correlation between items is expected to approach to zero. With a view to meeting the requirements of the local independence assumption, where just a single proficiency is required for responding all items, these items are considered unidimensional (Nandakumar, 1994). The compliance with the unidimensionality assumption can provide evidence regarding the local independence assumption (Hambleton, Swaminathan, & Rogers, 1991; Lord, & Novick, 1968). Given the fact that the present study meets the requirements of the unidimensionality assumption, it is also deemed to have met the requirements of the local independence assumption.

Once the assumptions were tested in accordance with the Item Response Theory, model-data fit was checked in order to identify the model offering the highest level of fit with the data set. The fit statistics calculated through separate calibrations for each grade revealed a state of affairs wherein

the 1 Parameter Logistics Model (PLM) and 2 PLM had model-data fit, while no model-data fit was observed for 3 PLM. Therefore, the analyses were applied in line with 2 PLM model.

FINDINGS and INTERPRETATION

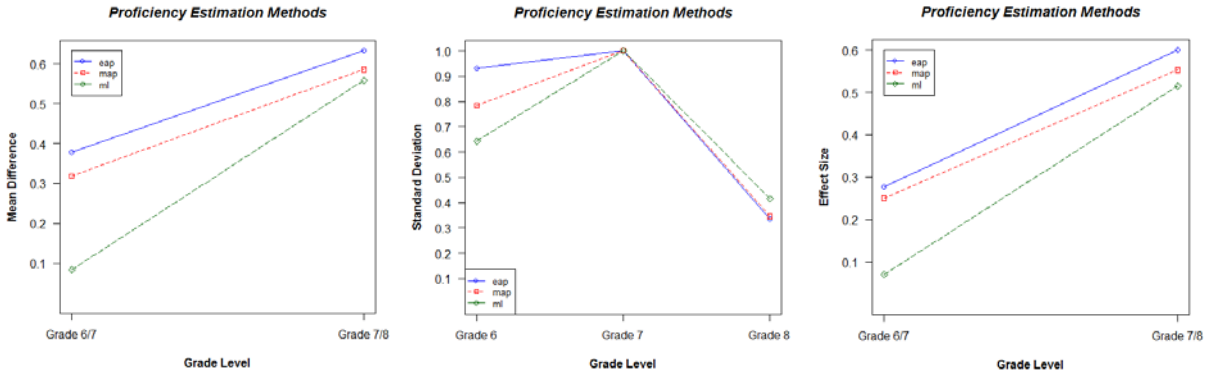
The findings of the study and the results obtained with reference to grade levels, calibration methods, and proficiency estimation methods employed were reviewed in light of mean, standard deviation, and effect size criteria.

In order to come up with an answer to first sub-problem, data pertaining to all grade levels were compiled in a single file, and all data were calibrated concurrently, using the software BILOG-MG 3. Concurrent calibration method was applied to estimate the item and proficiency parameters for each grade. The θ proficiency level means, mean differences, standard deviations and effect size values were established on the basis of ML, EAP and MAP proficiency estimation methods. The values thus calculated are presented below, in Table 2.

Table 2. Results of ML, EAP, and MAP Proficiency Estimation Obtained for Science Test through Concurrent Calibration Method

	Grade	ML	EAP	MAP
Mean	6	-0.084	-0.379	-0.318
	7	0.000	0.000	0.000
	8	0.558	0.633	0.585
Mean difference	7-6	0.084	0.379	0.318
	8-7	0.558	0.633	0.585
Standard deviation	6	0.643	0.930	0.785
	7	1.000	1.000	1.000
	8	0.415	0.336	0.346
Effect size	7-6	0.0709	0.2777	0.2505
	8-7	0.5154	0.6000	0.5530

Table 2 presents the evaluation criteria values for each grade. The graphs pertaining to these values are shown below, in Graph 1.



(a) (b) (c)

Graph 1. Graphs of Values Obtained Through the Concurrent Calibration Method: (a) Mean Differences, (b) Standard Deviations, (c) Effect Size.

As shown in both Table 2 and Graph 1 reveals, the means calculated through concurrent calibration on the basis of the data from the science test suggest that the proficiency level of the students increase as they progress from grade 6th to 8th. The review of mean differences with a view to ascertaining the criteria of development between individual grades suggests that the highest mean difference figures were observed with EAP, while the lowest ones were achieved with ML method.

The review of standard deviations, to assess the variability criteria between individual grades, on the other hand, reveals that the standard deviation fell as one moved from 6th grade to 8th, and the highest standard deviation was observed with EAP, while ML produced the lowest ones. As 7th grade was chosen as the reference year, all estimation methods stipulated a standard deviation of one (1) for that grade.

The analysis of effect sizes, with a view to evaluate the differentiation criteria between level distributions, reveals that effect size grew from 6th grade to 8th, with the largest effect sizes were observed with EAP, while the lowest ones were obtained with ML method. An analysis of the figures in Table 2 reveals that the effect size changes between the 6th and 7th grade can be considered small, while the one between the 7th and 8th is medium.

These findings run in parallel to the studies by Tong and Kolen (2010) and Kim (2007), using concurrent calibration method. Furthermore, the studies by Meng, Kolen and Lohman (2006) and Tong (2005) also found, in a similar vein, that the smallest effect size value was obtained through ML estimation.

In order to come up with an answer to second sub-problem, data for each grade level were calibrated separately using 2PLM. Item and proficiency parameters were calculated using BILOG-MG 3 software. In order to present the parameter estimations for each grade on the scale for the 7th grade, which is accepted as the reference level, the ST (Hanson, Zeng, & Chien, 2004) software, which is calculating IRT scaling constants and written in C programming language, was used. And also, Stocking Lord method was used to estimate the gradient and intersection values as a characteristics curve method.

Quadrature points are used for conversions applying Stocking Lord method. The analyses required for the calculation of Quadrature points were affected using the icl_win software. The quadrature points established thus were added to codes, to come up with SL conversion.

SL method was applied using the test-characteristic curves. The slope and intersection values produced are presented below in Table 3.

Table 3. Constants A and B calculated for Stocking Lord Conversion

Grade	A (Slope)	B (Intercept)
6-7	1.121	0.767
7-8	1.574	-0.962

The conversions are effected using the constants A and B obtained through the SL conversion presented in Table 3. Since 7th grade is set as the reference level, when converting the 6th grade to the 7th, proficiency estimations are effected through the equation "$\theta_{new}=\theta_{old}$ x 1.121 + (0.767)". On the other hand, conversion of the 8th grade to the 7th is done through the equation $\theta_{new}=\theta_{old}$ x 1.574 + (-0.962). A two-step conversion is required for transition from the 8th grade to the 6th. The equation $\theta_{new}=(\theta_{old}$ x 1.121 + (0.767)) x 1.574 + (-0.962) was used for the conversion of the 8th grade. The intersection values between the 6th and the 7th grades are positive, while those between the 7th and the 8th are negative.

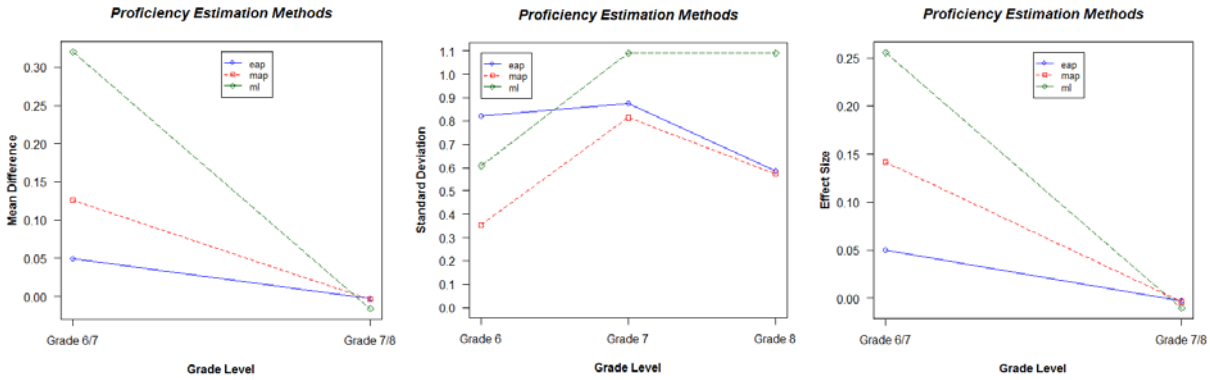
Estimation was effected using separate calibration method with the BILOG-MG 3 software using the calculated estimation values as well. The θ proficiency level means, mean differences, standard

deviations and effect size values were established on the basis of ML, EAP and MAP proficiency estimation methods. The results are presented below in Table 4.

Table 4. The Results of ML, EAP, and MAP Proficiency Estimations Obtained for Science Test through Separate Calibration Method

	Grade	ML	EAP	MAP
Means	6	-0.317	-0.058	-0.147
	7	0.007	0.002	-0.021
	8	-0.009	-0.005	-0.025
Mean difference	6-7	0.320	0.060	0.126
	7-8	-0.016	-0.007	-0.004
Standard deviation	6	0.608	0.822	0.354
	7	1.091	0.874	0.814
	8	1.091	0.586	0.575
Effect size	6-7	0.2558	0.0498	0.1420
	7-8	-0.0104	-0.0030	-0.0040

Table 4 presents the evaluation criteria values for each grade. To present a clearer picture of these figures, the graphs pertaining to these values are shown below, in Graph 2.



(a) (b) (c)

Graph 2. Graphs of Values Obtained Through the Separate Calibration Method: (a) Mean Differences, (b) Standard Deviations, (c) Effect Sizes.

As seen in both Table 4 and Graph 2 reveals, the means calculated through separate calibration on the basis of the data from the science test suggest that the proficiency level of the students increase as they progress from grade 6th to 7th, and fall from grade 7th to 8th. Mean differences, which reflect the level of improvement from one grade to another allows a better understanding of this criterion. While the mean differences are positive between grades 6th and 7th, they are negative between grades 7th and 8th, and tend to fall from grade 6th to 8th. This finding can be interpreted as the fact that the 7th grade students are more successful than the 8th grade students and that the desired and expected growth from one class level to the other class level cannot be achieved. The reason for the 8th grade students being less successful than the 7th grade may be the TEOG (Basic Education to Secondary Transition) exam. The increase in students' anxiety levels may have adversely affected their success. In addition, the fact that eighth grade students have entered adolescence may have affected their psychology and achievements negatively. In the study of Briggs, Weeks and Wiley (2009), parallel to this finding, it was stated that the growth patterns did not show an increase from one year to the

other year as a linear. It seems that there are studies supporting this finding in the literature (Tong, & Kolen, 2008; Cetin, 2009; Wysel, & Reckase, 2011; Altun, 2013). In Tong and Kolen (2010)'s study, it was found that the mean difference was higher in the lower class levels and the mean difference decreased as the class level increased. Similar to the results of Tong and Kolen's (2010) study, Ito, Skykes and Yao (2008)'s and Tong and Kolen (2007)'s studies, compared vertical scaling methods, have stated that he increase in the scores of the students in the lower grade level is higher than the increase in the scores of the students in the higher grade level. As a result of the IRT analyzes the scores of the students increase and decrease according to the grade levels. In other words, the success levels of unsuccessful students are increasing in 6th grade to 7th grade, compared to the transition from 7th grade to 8th grade. And, when the estimation methods are compared, it is seen that the highest mean differences were obtained with ML, while EAP produced the lowest ones.

A glance at standard deviation figures shows that overall standard deviation between grades 6th and 8th tend to fall. While the lowest standard deviation is established with ML method, EAP produced the highest level of standard deviation.

The analysis of effect sizes indicates that in all three methods, effect sizes tend to fall towards grade 8th, with the largest effect sizes being observed when ML is applied, in contrast to the smallest ones are obtained through EAP. An analysis of the figures in Table 4 reveals that the effect size changes between the 6th and 7th grades as well as between the 7th and 8th grades can be interpreted as a weak effect. The review of the literature reveals that these findings run in parallel to those of Tong and Kolen (2007).

DISCUSSION and CONCLUSIONS

The objective of this study is to apply vertical scaling based on item response theory, leading to a comparison of calibration methods and proficiency estimation methods, and the developmental vertical scale scores calculated with reference to the mean, standard deviation, and effect size values.

The means calculated through concurrent calibration on the basis of the data from the science test showed that the proficiency level of the students increase as they progress from grade 6th to 8th. The mean differences for all three grades present a picture where largest differences are produced with EAP method. A glance at standard deviation figures shows that standard deviation between grades 6th and 8th tends to fall, and the lowest standard deviation value is established with ML method. Effect size picture suggests an increase from grade 6th to 8th, with the largest effect size values being produced with EAP method.

When the separate calibration method is applied as another calibration, the developmental scale scores present an increase in the means from grade 6th to 8th, while mean differences fall approaching from 6th to grade 8th. The highest mean difference was observed with EAP method. The mean differences generated through separate calibration were also notably lower than those generated through concurrent calibration. Standard deviation picture presents falling rates as one move from grade 6th towards 8th. The lowest standard deviation was observed with ML method. The standard deviation values calculated in separate calibration were generally lower than those produced through concurrent calibration. On the effect size front, it is observed that the effect sizes values decreasing from 6th grade to 8th grade. The highest effect size was observed with ML method. The effect size values calculated in separate calibration were lower than those produced through concurrent calibration.

The comparison of the findings pertaining to the first and the second sub-problems reveals that the mean differences found through separate calibration were lower than those applicable to concurrent calibration, while the standard deviation found in the case of separate calibration were again lower than the values established through concurrent calibration. Furthermore, the scale of impact in the case of separate calibration was again lower than the values applicable to concurrent calibration. The results reached for all three criteria, using the concurrent calibration method were ranked in the order ML < MAP < EAP, with ML leading to the lowest value while EAP producing the highest one. In

case of separate calibration, on the other hand, the ranking of results was found to vary with reference to the criteria applied.

The conclusions reached through the study reveal that vertical scaling is a complex process, and that there is no single all-applicable method. Since there is no single method supported by a wide-ranging consensus, taking into account the complexities of the methods applied and the results of the analyses, it is recommended that the researcher should decide on the method to apply, within the context of her specific study. The interactions between the issues discussed in this process can have an impact on the results of vertical scaling, and hence on the interpretations about the ongoing development of the students' achievements, one can recommend effective comparisons employing a range of methods, to lead to decisions regarding the achievements of students. Hanson and Béguin (2002) also emphasized that no single all-applicable method can be designated, and that comparing results through a combination of various equating methods under different conditions is the way to go.

Such an analysis should actually be considered an inherent part of the overall vertical scaling process. Test developers and users can be recommended to work on the process of equating the observed and actual scores in the final stage of the vertical scaling process, with the review of factors affecting observed scores.

Achievement levels of the students were observed to increase as one move from earlier grades to subsequent ones. However, further studies may be needed to assess whether such increases are at required levels or not. In order to ascertain the level of change students experience from one grade to another, vertical scaling practices are crucial. Vertical scaling assessments can be recommended to review the students' achievements at the K-12 level.

In the present study, test length (40 items), number of anchor items (10), sample size (1500), and applied model (2PLM) were fixed, and not subjected to analysis as determining factors or independent variables. Other studies can use these as variables in their own right, and investigate their impact on vertical scaling results as well. It is also possible to carry out a longitudinal study to review the achievement levels of individual students through extended years, followed up by an analysis on the basis of data from such longitudinal study. Since there is no single and exact criteria to assess the applicability of the methods employed in vertical scaling, the researchers are recommended to use more than one evaluation criteria (mean, mean differences, standard deviation, effect sizes, vertical distance, root-mean square error of approximation (RMSEA) and bias values) when comparing scaling results.

REFERENCES

Altun, A. (2013). *Dikey ölçeklemede madde tepki kuramına dayalı farklı kalibrasyon ve yetenek kestirim yöntemlerinin karşılaştırılması* (Unpublished Doctoral Thesis). Ankara: Hacettepe University.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008, April). *Vertical scaling in value-added models for student learning*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Boughton, K. A., Lorie, W., & Yao, L. (2005). *A multidimensional multi-group IRT models for vertical scales with complex test structure: An empirical evaluation of student growth using real data*. Paper presented at the annual meeting of the National Council on Measurement in Education, Monreal, Canada.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative and mixed methods approaches* (4th edition). University of Nebraska, Lincoln: Sage.

Cetin, E. (2009). *Dikey ölçeklemede klasik test ve madde tepki kuramına dayalı yöntemlerin karşılaştırılması* (Unpublished Doctoral Thesis). Ankara: Hacettepe University.

Dongyang, L. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design* (Unpublished Doctoral Thesis). University of Maryland.

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

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

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*, 3-24.

Hanson, B. A., Zeng, L., & Chien, Y. (2004). *ST: A computer program for IRT scale transformation* [Computer software]. Retrieved January 24, 2005, from http://www.education.uiowa.edu/casma.

Harris, D. J. (2003). Equating the multistate bar examination. *The Bar Examiner, 72*(3), 12-18.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (pp. 187–220). Westport, CT: Praeger.

Ito, K., Sykes, R. C., & Yao, L. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurment in Education, 21,* 187-206.

Karkee, T. B. & Wright, K. R. (2004). *Evaluation of linking methods for placing three-parameter logistic item parameter estimates onto a one-parameter scale*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Kim, J. (2007). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales* (Unpublished Doctoral Thesis). University of Iowa.

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kim, J., Lee, W. C., Kim, D., & Kelley, K. (2009). *Investigation of vertical scaling using the Rasch model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd Ed.) New York: Springer Verlag.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-193.

McBridge, J., & Wise, L. (2001). *Developing the vertical scale for the Florida comprehensive assessment test (FCAT).* Paper presented at the annual meeting of the Harcourt Educational Measurement, San Antonio, Texas.

Meng, H (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.* (Unpublished Doctoral Thesis). University of Iowa, Iowa.

Meng, H., Kolen, M. J., & Lohman, D. (2006). *An empirical investigation of IRT scaling methods: How different IRT models, parameter estimation procedures, proficiency estimation methods, and estimation programs affect the results of vertical scaling for the cognitive abilities test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17-35.

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

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44*(3), 249-275.

Tong, T. (2005). *Comparison of methodologies and results in vertical scaling for educational achievements tests* (Unpublished Doctoral Thesis). University of Iowa, Iowa.

Tong, Y., & Kolen, M. (2007). Comparison of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.

Tong, Y., & Kolen, M. (2008, March). *Maintenance of vertical scales.* Paper presented at the National Council on Measurement in Education, New York City.

Tong, Y., & Kolen, M. (2010). Scaling: An ITEMS module. *Educational Measurement: Issues and Practice, 29*(4), 39-48

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating.* New York: Springer.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of Item Response Theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*(1), 11-26.

Wysel, A. E., & Reckase, M. D. (2011). A graphical approach to evaluating equating using test characteristic curves. *Applied Psychological Measurement, 35*(3) 217–234.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93-111.

Zhu, W. (1998). Test equating: What, why, who? *Research Quarterly for Exercise and Sport, 69*(1), 11–23.

UZUN ÖZET

Giriş

Dikey ölçekleme sürecinde farklı veri toplama desenleri, ölçekleme yöntemleri, kalibrasyon yöntemleri, yetenek kestirimi yöntemleri ve değerlendirme ölçütleri kullanılabilir. Araştırmacıların ölçekleme sürecinde kullanılacak desen ve yöntemlere ilişkin çeşitli kararlar vermesi gerekmektedir. Bu kararların dikey ölçeklemeyi dolayısıyla da öğrenci başarısındaki gelişimi gösteren örüntüleri etkilediği görülmüştür (Tong & Kolen, 2007). Bu çalışmada veri toplama deseni olarak denk olmayan gruplarda ortak madde deseni, ölçekleme deseni olarak Madde Tepki kuramına dayalı 2 Parametreli lojistik model kullanılmıştır. Sınıf seviyelerinin ortak bir ölçeğe bağlanması için kullanılan ölçek dönüştürme kalibrasyon yöntemlerinden ayrı ve eş zamanlı kalibrasyon; madde parametrelerini kestirebilmek için kullanılan kestirim yöntemlerinden ise, Maximum Likelihood Estimation (ML) (Maksimum Olabilirlik), Expected A Posteriori (EAP) (Beklenen Önsel Dağılım) ve Maximum A Posteriori (MAP) (Maksimum Önsel Dağılım) kestirim yöntemleri kullanılmıştır. Ölçekleme çalışmasının son aşamasında ise elde edilen sonuçlar bir sınıf düzeyinden diğer sınıf düzeyine olan büyüme, sınıf düzeyleri arasındaki çeşitlilik ve düzey dağılımlarının ayrımı değerlendirme ölçütleri kullanılarak karşılaştırılmıştır.

Alan yazın incelendiğinde gerçek veri ile yapılan çalışmaların oldukça az olduğu, daha çok simülasyon verileri ile yapılan çalışmalara ağırlık verildiği görülmektedir. Bu çalışmada araştırmacılar tarafından geliştirilen fen bilgisi başarı testi 1500 öğrenciye uygulanarak toplanan gerçek veriler üzerinde analizler yürütülmüştür, böylece bu çalışmanın alan yazına katkı sağlayacağı düşünülmektedir.

Yöntem

Araştırmada var olan yöntem ve teknikler gerçek veri ve yapay veri üzerinden sınandığı ve en az hatalı yöntemler belirlenerek kuramsal çalışmalara katkı sağlaması amacı taşıdığı için araştırma temel araştırma niteliğindedir (Creswell, 2013). Araştırmada çalışma grubu 6ncı, 7nci ve 8inci sınıf öğrencilerinden oluşmaktadır. Çalışma grubu, Ankara ili Altındağ, Çankaya, Gölbaşı, Keçiören ve Mamak ilçelerinden ikişer okul olmak üzere 12 farklı okuldan toplam 1500 öğrenciden oluşmaktadır. Uygulanan fen bilgisi başarı testi için 2008-2012 yılları arasında uygulanan SBS (Seviye Belirleme Sınavı), OKS (Ortaöğretim Kurumları Seçme ve Yerleştirme Sınavı) ve PYBS (Parasız Yatılılık ve Bursluluk Sınavı) testlerinden ayırt edicilik düzeyleri ve madde güçlük indeksleri kontrol edilerek maddeler seçilmiş ve üç sınıf düzeyine uygun 40ar maddelik birer başarı testi geliştirilmiştir. Bu testlerde ardışık sınıflar arası zincirleme ölçeklemeyi sağlayacak 10'ar madde ortak madde olarak belirlenmiştir. Hambleton, Swaminathan ve Rogers (1991), ortak maddelerin sayısının testin tamamının %20'si kadar olmasının uygun olduğunu belirtirken, birçok araştırmada ortak madde sayısındaki artışın testteki ölçmenin standart hatasını azalttığını belirtilmektedir (Boughton, Lorie & Yao, 2005; Kim, Lee, Kim & Kelley, 2009). Bu nedenle bu çalışmada toplam madde sayısının %25'i kadar ortak madde kullanılmıştır. Bu araştırmada denk olmayan gruplarda ortak madde deseni kullanılmıştır. Bu desen uygulamada yaygın olarak kullanılan desenlerden biri olmakla birlikte, en esnek ve en karmaşık desenlerden biridir (Sinharay & Holland, 2007). Pratiklik açısından tercih edilen bir yöntem olmakla birlikte, diğer desenlere göre de daha az sınırlayıcıdır (Zhu, 1998).

Sonuçlar ve Tartışma

Analizler yapılmadan önce veri temizleme yapılarak eksik ve kayıp veriler veri setinden çıkarılmış ve fen bilgisi başarı testinden elde edilen puanların MTK varsayımlarından tek boyutluluk, yerel bağımsızlık ve model veri uyumu kontrol edilmiştir. Birinci ve ikinci alt probleme ait bulgular incelendiğinde; dikey ölçekleme analizinde farklı kalibrasyon yöntemlerinden elde edilen sonuçlar

karşılaştırıldığında; ayrı kalibrasyon ile elde edilen ortalama farkların eş zamanlı kalibrasyon ile elde edilen ortalama farklarından daha düşük olduğu, ayrı kalibrasyon ile elde edilen standart sapma değerlerinin genel olarak eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu ve ayrı kalibrasyon ile elde edilen etki büyüklüğü değerlerinin eş zamanlı kalibrasyon ile elde edilen değerlere göre daha düşük olduğu görülmektedir. Eş zamanlı kalibrasyon yöntemi ile her üç kriter için de elde edilen sonuçların ML < MAP < EAP şeklinde sıralandığı; en küçük değerlerin ML, en büyük değerlerin ise EAP ile elde edildiği görülmektedir. Ayrı kalibrasyon da ise sonuçların sıralamalarının kriterlere göre değiştiği görülmektedir. Araştırma bulgularına göre, dikey ölçekleme sürecinin karmaşık bir süreç olduğu ve tek bir doğru yöntem olmadığı görülmektedir. Üzerinde hemfikir olunan doğru bir yöntem olmadığı için, uygulanan yöntemlerin karmaşıklığı analizlerin sonuçları göz önünde bulundurularak en uygun yöntemi yine araştırmacı araştırmasına uygun olarak belirleyebilir. Bu süreçte ele alınan koşulların birbiriyle etkileşimi dikey ölçekleme sonucunu dolayısıyla öğrenci başarısının gelişimine yönelik yapılacak yorumları etkileyebileceği için öğrenci başarıları hakkında karar verirken farklı yöntemlerin de kullanılarak karşılaştırma yapılması önerilebilir. Hanson ve Béguin (2002) de tek bir doğru yöntem belirtilemeyeceği, farklı koşullarda doğru yöntemi belirleyebilmek için eşitleme yöntemlerini bir arada kullanarak, sonuçlarını karşılaştırmanın etkili olacağını vurgulamışlardır. Öğrenci başarılarının genel olarak ardışık sınıf seviyesi arttıkça arttığı görülmüştür, fakat bu artışın istendik düzeyde olup olmadığını değerlendirebilmek için çalışmalar yapılabilir. Öğrencilerin yıldan yıla başarılarındaki değişimin belirlenebilmesi için dikey ölçekleme uygulamaları oldukça önemlidir. Öğrencilerin K-12 seviyesinde başarılarının takibi için dikey ölçekleme çalışmalarının başlatılması ve yürütülmesi önerilebilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);112-127

WÇZÖ-IV Maddelerinin Cinsiyet ve Sosyo-Ekonomik Düzey Açısından İşlev Farklılığının Belirlenmesinde Kullanılan Yöntemlerin Karşılaştırılması*

Gender and Socioeconomic Status DIF on The WISC-IV Turkish Form Items: A Comparison of DIF Detection Tecniques

Elif Bengi ÜNSAL ÖZBERK ** Nizamettin KOÇ ***

Öz

Bu araştırmanın amacı Wechsler Çocuklar için Zekâ Ölçeği (WÇZÖ) IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre işlev farklılığı gösterip göstermediğinin, birden fazla yöntemle incelenerek birbiri ile uyumlu sonuçlar verip vermediğinin ortaya konulmasıdır. Araştırmada WÇZÖ-IV Türkiye Uyarlama ve Standardizasyon Çalısmasına ait 819 kişilik ön uygulama verileri kullanılmıştır. Çalışma WÇZÖ-IV'ün çoklu puanlanan maddelerden oluşan Küplerle Desen, Benzerlikler, Sayı Dizisi, Sözcük Dağarcığı, Harf-Rakam Dizisi, Kavrama alttestleri ve ikili puanlanan maddelerden oluşan Resim Kavramları, Mantık Yürütme Kareleri, Resim Tamamlama, Genel Bilgi, Aritmetik, Sözcük Bulma olmak üzere toplam 12 alttest ve bu alt testlerde yer alan 315 madde üzerinden yürütülmüştür. Maddelerin işlev farklılığı içerip içermediğine karar vermek için ikili puanlanan maddeler açısından Rasch Modeli, Mantel-Haenszel, SIBTEST yöntemleri, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli, Mantel Test ve Poly-SIBTEST yöntemleri kullanılmış ve 12 alt test açısından madde işlev fonksiyonu içeren maddeler belirlenmiştir. Araştırmada kullanılan MİF belirleme yöntemleri incelendiğinde, genel olarak yöntemlerin birbiriyle tutarlı sonuçlar verdiği ancak cinsiyet açısında MİF içeren maddelerin tespitinde Mantel-Haenszel, SIBTEST ve Mantel Test, Poly-SIBTEST istatistikleri daha tutarlı sonuçlar verirken sosyoekonomik düzeye göre MİF içeren maddelerin tespitinde Mantel-Haenszel, Rasch Modeli ve Mantel Test, Kısmi Puan Modelinin daha tutarlı sonuçlar verdiği saptanmıştır.

Anahtar Kelimeler: Wechsler çocuklar için zekâ ölçeği IV, Rasch modeli, Mantel-Haenszel, SIBTEST, kısmi puan modeli, Mantel test, Poly-SIBTEST.

Abstract

The purpose of this study is to investigate potential gender and socio-economic status bias in theWechler Intelligence Scale for Children: Fourth Edition (WISC-4) by using several differential item functioning detection techniques. In this study, WISC-4 Turkish standardization test pilot data including 817 children were used. In accordance with the purpose of the study, 315 items were used both in polytomously scored subtests such as Block Design, Similarities, Digit Span, Vocabulary, Letter-Number Sequencing, Comprehension, and dichotomously scored subtests such as Picture Concepts, Matrix Reasoning, Picture Completion, Information, Arithmetic, and Word Reasoning. While Rasch Model, Mantel-Haenszel, and SIBTEST DIF detection techniques were used for dichotomously scored items, Partila Credit Model, Mantel, and Poly-SIBTEST techniques were used for polytomously scored items. In terms of DIF techniques, Mantel-Haenszel, SIBTEST and Mantel Test, Poly-SIBTEST analyses provided similar results when DIF based on gender was investigated. In addition Mantel-Haenszel, Rasch estimations and Partial Credit Model, Mantel Test results were similar while investigating DIF according to socioeconomic status.

Keywords: Wechsler intelligence scale for children IV, rasch model, Mantel-Haenszel, SIBTEST, partial credit model, Mantel test, Poly-SIBTEST.

* Bu çalışma, birinci yazarın Prof. Dr. Nizamettin KOÇ danışmanlığında tamamlanan doktora tezinden türetilmiştir.

** Dr., Adalet Bakanlığı Ceza ve Tevkifevleri Genel Müdürlüğü, Personel Eğitim Bürosu, Şube Müdürü, Ankara-Türkiye, elifbengiunsal@gmail.com
*** Prof. Dr., Ankara Üniversitesi, Eğitimde Ölçme ve Değerlendirme ABD, Ankara-Türkiye, Emekli, nkoc@ankara.edu.tr

DOI: *10.21031/epod.287577*
Geliş Tarihi: 24.01.2017
Kabul Tarihi: 02.03.2017

GİRİŞ

Psikolojik testlerin genel amacı; kişiler arasındaki farkları belirleyip karşılaştırabilmek (bireylerarası-interindividual) ya da aynı bireyin zaman içerisinde değişen (bireyiçi-intraindividual) yeteneğini ölçebilmektir (Cronbach, 1990). Bu yüzden psikolojik testlerdeki değerlendirmenin amacı bireylerarası ve bireyiçi farklılıkları belirlemektir. Bu noktada eğitsel ve psikolojik ölçmeler ve buna bağlı olarak ölçme araçlarının geliştirilmesi, bireyler arası ve birey içi fark kavramından doğmuştur (Thorndike, 1982). Test sonuçlarının testi alan alt gruplara göre değişiklik göstermesi, gruplar arasındaki gerçek yetenek veya başarı farklılığından kaynaklanmış olabileceği gibi testin sebep olduğu sistematik bir adaletsizliğin sonucu da olabilir.

Cronbach (1990)'a göre testlerin kullanım amaçları, bireylerin belli bir özelliğe göre sıralanması, seçilmesi, bir üst sınıf ya da kademeye geçirilmesi kararının verilmesi olarak sıralanmıştır. Murphy ve Davidshofer (1994)'in belirttiği gibi testler bireyler hakkında önemli kararlar almak için kullanılmaktadırlar. Benzer bir şekilde, Türkiye'de özel eğitime ihtiyacı olan çocukların belirlenmesi ve özel eğitime devam kararlarının verilmesinde ayrıca üstün yetenekli çocukların belirlenmesi ve Bilim Sanat Merkezleri'nde eğitim almaları gibi önemli kararlarının verilmesinde Wechsler Çocuklar için Zeka Ölçeği kullanılmaktadır.

Böylesine önemli kararların verildiği, testlerin sonuçlarına bağlı olarak adil kararlar verebilmek için testlerin psikometrik özelliklerinin kabul edilebilir seviyelerde olması gerekmektedir (Horst 1966; Reynold, Livingston ve Willson, 2006). Bu amaçla, testlerden beklenen de hatalardan olabildiğince arınık olması ve ölçmek istediği özelliği başka değişkenler karışmadan ölçebilmesidir. Bir testin ölçmek istediği özellik dışında başka değişkenlerin karışmasına örnek oluşturan durumlardan biri de testin yanlılık içermesidir. Yanlılık, farklı alt gruplardaki bireylerin test puanlarının bulundukları gruba bağlı olarak sistematik hata içermesidir (Camilli ve Shepard, 1994; Tittle, 1988; Zumbo, 1999). Test yanlılığı, bir testin sonuçlarına dayanılarak alınan bir karar, tüm gruplar için adil değilse ya da bir gruba, diğerine göre eşit olmayan bir etki yapıyorsa ortaya çıkar (Osterlind ve Everson, 2009). Eğer bir test yanlı ise, testi alan farklı gruplardaki bireylere adil davranılmamasına neden olmaktadır (Tittle, 1988). Madde yanlılığı ise birçok araştırmacı tarafından, aynı yetenek düzeyinde olan, fakat cinsiyet, sosyoekonomik düzey, etnik köken, vb. gibi farklı gruplardan gelen bireylerin, test koşullarından ya da maddenin bazı özelliklerinden, test maddelerine doğru cevap verme olasılıklarının diğer gruba göre az ya da çok olması biçiminde tanımlanmıştır (Adams ve Rowe, 1988; Mellenberg, 1983; Osterlind, 1983; Raju, 1990; Rodney ve Drasgow, 1990; Shepard, Camilli ve Williams, 1984; Tittle, 1988; Zumbo, 1999).

Madde yanlılığı çalışmaları özellikle test geliştirme ve uyarlama sürecine önemli katkı sağlar. Özellikle testlerin denkliğini sağlama açısından gerekli bir süreçtir. Madde yanlılığı çalışması ile her bir maddeler gözden geçirilerek, her bir madde için ayrı ayrı geçerlik kanıtı toplanmış olur. Madde yanlılığı çalışmaları, hem istatistiksel hem de uzman kanılarına dayalı yargısal süreçleri gerektirir. Test maddelerinin yanlı olup olmadığının belirlenmesinin ilk adımı madde işlevinin farklılığını (MİF) belirlemeye yönelik istatistiksel bir süreçtir. Belirli istatistiksel işlemler sonrası MİF gösteren bir madde olası yanlı madde olarak değerlendirilir (Kamata ve Vaughn, 2004). Bu açıdan madde yanlılığı çalışmalarında, madde etkisi ve MİF kavramları arasındaki farka değinmek önemlidir. Madde etkisi, farklı alt gruplardan gelen bireylerin bir maddeyi doğru yanıtlama olasılıklarının, ilgili madde ile ölçülmek istenen psikolojik özellik bakımından farklılaşmasını ifade etmektedir (Zumbo, 1999). Bu farklılık, madde etkisi söz konusu olduğunda, madde ile ölçülen psikolojik özellik bakımından gruplar arasındaki gerçek farklardan kaynaklanırken, MİF söz konusu olduğunda grupların yetenek düzeylerindeki farklılaşmadan değil, MİF'ten kaynaklanmaktadır (Camilli ve Shepard, 1994). Bir maddenin yanlılık içermesi, maddenin MİF içerdiğinin göstergesidir. Ancak MİF gösteren maddelerin yanlı olduğu kesin değildir (Kamata ve Vaughn, 2004). Bir maddenin yanlı olduğu kararını vermek için o maddenin yanlızca MİF göstermesi yeterli olmamakla birlikte maddenin MİF göstermesi, maddenin yanlılığının ortaya konması sürecinde bir ilk adımdır.

MİF belirleme çalışmalarının başladığı günden bugüne birçok MİF belirleme yöntemi kullanılmıştır. Örneğin Zumbo (2007) sınıflandırmayı olasılık tablosuna dayalı yöntemler, madde tepki kuramına dayalı yöntemler, çok boyutlu yöntemler biçiminde yaparken Camilli, Shepard (1994) varyans analizine dayalı yöntemler, olasılık tablosuna dayalı yöntemler olarak yapmıştır. MİF çalışmalarında birden çok yönteme dayalı inceleme yapılması önerilmektedir (Holland ve Wainer, 1993; Osterlind ve Everson, 2009). Bu araştırmada da, cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddeler, Wechsler Çocuklar için Zeka Ölçeği IV Türkçe formundaki ikili puanlanan maddeler açısından Rasch MODELİ, SIBTEST ve Mantel-Haenszel, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli, Poly-SIBTEST ve Mantel Test ile yapılan analizlerle incelenmiştir.

İkili puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntemlerden biri olan Rasch modeli, Rasch tarafından geliştirilen MTK kapsamında tek parametreli bir modeldir (Rasch, 1960). Rasch modelinde sadece güçlük parametresi (bi) kullanılır ve model ayırıcılık parametresini tüm maddeler için sabit tutar, şans parametresi ise 0'dır. Rasch modelinde madde karakteristik eğrisi aşağıdaki fonksiyonla bulunmaktadır (Hambleton, Swaminathan ve Rogers, 1991).

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)} \quad (1)$$

θ_S bireyin yetenek düzeyini β_i, i maddesinin güçlük indeksini ifade etmektedir. $P(X_{IS}=1|\ \theta_S, \beta_i)$de, β_i güçlük düzeyindeki maddeyi, θ_S yetenek düzeyindeki bireyin doğru cevaplama olasılığını tanımlamaktadır (Embretson ve Reise, 2000). İkili puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan bir diğer yöntem olan Mantel Haenszel yöntemi, Mantel ve Haenszel (1959) tarafından eşleştirilmiş gruplarda uygulanacak olan ki-kare tekniği olarak geliştirilmiştir. Bu yöntem, 1998 yılında Holland ve Thayer tarafından güncellenerek MİF belirleme yöntemi olarak ölçme değerlendirme literatürüne kazandırılmıştır. MH yöntemi χ^2 istatistiğine dayanır, MH yöntemi için odak (focal) ve referans gruplarındaki bireyler gösterdikleri performansa göre eşleştirilir. Bu eşleştirmenin yapılabilmesi için referans ve odak grupta yer alan cevaplayıcıların testten aldıkları toplam puanlara göre 4 ya da 5 yetenek grubu oluşturulur. Heterojen puan dağılımlarında yetenek grubu sayısı artırılabilir. Odak ve referans grupların toplam puanları eşleştirildikten sonra her madde için 2 (gruplar) x 2 (madde puanları) x M (puan düzeyi) olasılık çizelgesi olarak isimlendirien üç boyutlu bir matris oluşturulur. MH_{χ}^2, 1 serbestlik derecesinde, yaklaşık normal dağılıma sahip bir χ^2 istatistiğidir. MH_{χ}^2 istatistiği için kritik değer 0.05 manidarlık düzeyinde 3.84, 0.01 manidarlık düzeyinde ise 6.63'tür (Penfield, 2013). Son olarak bu çalışmada ikili puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntem olan SIBTEST, bir veya birde çok maddede MİF değerlendirmede kullanılan ve hipotez testi yöntemini kullanan nonparametrik bir yöntemdir. Bu yöntem Shealy ve Stout (1993)'ün çok boyutlu madde tepki kuramına göre MİF belirleme yöntemlerine dayanmaktadır. Model daha sonra Roussos ve Stout (1996) tarafından geliştirilmiştir. SIBTEST, testin ölçmek istediği örtük özellik bakımında bireyleri birbirleri ile eşleştirir ve bir veya birden çok madde üzerindeki performanslarda bir farklılık olup olmadığını inceler. SIB istatistiği, β_U, hesaplanırken ilk başta N maddelik test n maddelik alt testlere ayrılır ve N-n kadar şüpheli madde içerir. SIBTEST istatistiği yanlı madde olmadığı durumda N(1,0) olacak biçimde normal dağılıma sahiptir ($\beta_U = 0$) Bu yüzden odak gruba karşı MİF hipotezi H_0: $\beta_U = 0$ ve H_A: $\beta_U > 0$ ile test edilebilir.

Çoklu puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntemlerden biri olan Kısmi puan modeli, Rasch modelinin, çoklu puanlanan maddeler için uyarlanmış bir uzantısı olarak düşünülebilir. Kısmi puan modelinde, Rasch modelindeki maddenin güçlük parametresinin yerini,δ_{ij} , madde adım güçlüğü (item step difficulty) parametresi almaktadır. δ_{ij} , ardışık kategori yanıt eğrilerinin kesiştiği noktanın yetenek düzeyi

olarak yorumlanabilir. Bu yüzden madde parametreleri kategori kesişim parametresi (category intersection parameter) olarak da adlandırılmaktadır (Embretson ve Reise, 2000). δ_{ij} değeri ne kadar büyürse, bir maddeyi yanıtlamak için gerekli adımlardan ilgili olanının o kadar güç olduğu şeklinde yorumlanır. Kısmi puan modeli için herhangi bir kategoride yanıt verme olasılığı aşağıdaki formülle gösterilir:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{j=0}^{x}(\theta-\delta_{ij})\right]}{\sum_{r=0}^{mi}\left[\exp\sum_{j=0}^{r}(\theta-\delta_{ij})\right]} \quad (2)$$

Formulde mi; kategori eşik paremetre sayısını, δ_{ij}; j ile puanlanan kategoriye ait adım güçlüğü parametresini, x tepki kategorilerini temsil etmektedir. Çoklu puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntemlerden bir diğeri olan Mantel test, 1993 yılında ise Zwick ve diğerleri tarafından Mantel-Haenszel Testinin bir uzantısı olarak çok kategorili puanlanan maddelerde MİF belirlemede kullanılması önerilmiştir. MH χ^2 yöntemi çoklu puanlanan maddelere uyarlanmak istenildiğinde madde yanıt kategorilerinin sıralı ve gruplar arasında karşılaştırılabilir olduğu varsayımı öne çıkar. Mantel testi, hedef gruplardaki puanların eşleştirilmesine bağlı olarak madde ile grup üyelikleri arasındaki ilişkiyi inceler. Burada bahsedilen puan maddelerin toplamından elde edilen gözlenen puandır. Mantel test, MİF'e ilişkin yokluk hipotezini test etmek üzere bir serbestlik derecesinde ki-kare dağılımına ilişkin istatistik vermektedir (Zwick, Donoghue ve Grima, 1993). Son olarak bu çalışmada çoklu puanlanan maddelerde cinsiyete ve sosyoekonomik duruma göre MİF gösteren maddelerin tespiti açısından kullanılan yöntem olan Poly-SIBTEST'te ise, Shealy ve Stout (1993) tarafından geliştirilen SIBTEST ikili puanlanan maddelerde MİF analizi için düzenlenen formunun, Chang ve arkadaşları (1996) tarafından yapılan araştırmada ikili puanlanan maddelerin dışında benzer prosedür sıralama yanıtlarına dayanan çok kategorili maddeler için uygulanmış halidir.

Geçmişten bugüne madde işlev farklılığını belirlemek için kullanılan araştırmalarda yöntemlerin belirli koşullar atında üstün yanları ve sınırlılıkları tartışılmaktadır. Örneğin Ackerman ve Evans'ın 1992 yılında yaptığı çalışmada çok sayıda maddenin MİF gösterdiği durumlarda SIBTEST'in MH yönteminden daha iyi performans gösterdiği sonucuna ulaşmıştır. Yine aynı koşul altında Roussoss (1992) SIBTEST'in madde işlev farklılığını belirlemede I. tip hata oranının MH yönteminden daha kabul edilebilir olduğunu göstermiştir. Bir diğer koşul olan benzer yetenek dağılımına sahip büyük ve küçük örneklemler için madde işlev farklılığını belirlemede I. tip hata oranının MH ve SIBTEST için düşük fakat yetenek dağılımları arasındaki fark arttığında I. tip hata oranının da yüksek olduğu sonucuna ulaşılmıştır (Roussos ve Stout, 1996; Shealy ve Stout 1993). Uttaro ve Millsap (1994) MH yönteminin farklı test uzunluklarında farklı I. tip hata oranları gösterdiğini bulmuşlardır. Chang ve Mazzeo (1993), poly-SIBTEST yöntemini geliştirmek amacıyla simülasyon veri üzerinden yaptıkları çalışmalarında poly-SIBTEST yöntemi ile GMH ve Mantel yöntemlerini karşılaştırmışlar ve poly-SIBTEST yönteminin MİF belirlemede Mantel Test ve GMH yöntemi kadar başarılı olduğu sonucuna ulaşmışlardır. Mellor ise 1995 yılında yaptığı çalışmasında çok kategorili maddelerde MİF belirlemek için GMH, poly- SIBTEST, OLR, LDFA yöntemlerini karşılaştırmış ve çalışmada 4 yöntemin farklı yetenek dağılımlarında tek biçimli olan ve olmayan MİF'i belirleme gücü ve yöntemlerin 1. tip hata oranlarını incelemiştir. Araştırma hem simülasyon veri hem de gerçek veri üzerinden yürütülmüş ve iki grubun yetenek dağılımlarının aynı olduğu durumlarda dört yöntemin de her kategoride tek boyutlu MİF'i belirlemede başarılı fakat GMH ve poly-SIBTEST'in diğer yöntemlerden daha hassas olduğu sonucuna ulaşmıştır. Roussos ve Stout (1996) similasyon veriyle yaptığı çalışmasında, küçük örneklemlerde MH ve SIBTEST yöntemlerinin I. tip hata oranları arasında büyük farklılıklar tespit etmemiştir. Fakat 3000 kişilik büyük örneklemde SIBTEST ve MH yöntemleri MİF içermeyen maddeler için orta ve önemli düzeyde MİF içerdiği yönünde isabetli olmayan kararlar vermiştir. Henderson (1999), çok kategorili verilerde MİF belirlemek için kullanılan GMH, poly-SIBTEST ve LDFA yöntemlerini gerçek veri kullanarak cinsiyet açısından karşılaştırmıştır. Araştırmanın sonucunda, Poly – SIBTEST yöntemi en çok sayıda MİF gösteren

maddeyi belirlemiştir. Ayrıca bu maddelerin büyük bir çoğunluğu GMH ve LDFA analizleri sonuçlarıyla da uyumlu bulunmuştur. Gierl, Jodoin ve Ackerman (2000)'nın farklı örneklem büyüklükleri ile yaptıkları çalışmada örneklem büyüklüğü arttıkça SIBTEST yöntemi ile MİF gösteren madde sayısının arttığı sonucuna ulaşmışlardır. Ayrıca bu çalışmaya ek olarak bir çok çalışmada SIBTEST için örneklem büyüklüğü arttıkça yöntemlerin I. tip hata oranının arttığı bulunmuştur (Narayanan ve Swaminathan, 1994; Rogers ve Swaminathan, 1993; Rousses ve Stout, 1996). Awuor, 2008 ise çalışmasında eşit olmayan örneklem grupları için MH yönteminin I. Tip hatayı SIBTEST yönteminden daha isabetli olarak kontrol ettiği sonucuna ulaşmıştır. Taylor ve Lee (2012) çalışmalarında cinsiyete ilişkin yanılılığı Poly-SIBTEST ve Rash modeli yöntemleri ile incelemişlerdir. Araştırmanın sonucunda Rasch modeli SIBTEST yönteminden daha çok maddeyi MİF'li olarak tespit etmiştir.

Atalay, Gök, Kelecioğlu ve Arsan (2012), örneklem büyüklüğü (odak ve referans gruplar için eşit örneklem büyüklükleri 400-400; 1500- 1500), yetenek dağılımı [(N(0,1) ve N(0,1); (N(0,1) ve N(0.5,1)] ve testteki MİF'li madde oranı (%5 ve %10)'nın değiştiği koşullarda SIBTEST, MTK-OO, MH ve LR yöntemlerini karşılaştırdıkları bir simülasyon çalışması yapmışlardır. Araştırma sonucunda, MİF'li maddeleri belirlemede MTK-OO yönteminin SIBTEST yönteminden; SIBTEST yönteminin LR ve MH yöntemlerinden, MH yönteminin ise LR yönteminden daha duyarlı olduğu sonucuna varılmıştır. Ayrıca MH yönteminin diğer yöntemlerle uyum yüzdesinin genel olarak düşük olduğu sonucuna ulaşmışlardır. Ulutaş (2012) PISA 2006 fen okuryazarlığı testinde yer alan maddelerin cinsiyet açısından MİF gösterip göstermediğini incelemiş, Türkiye ve Amerika arasında kültürlerarası eşdeğerliğini araştırmıştır. Araştırmacı MİF gösteren maddelerin belirlenmesi için klasik test kuramına dayanan Mantel-Haenszel, SIBTEST ve madde tepki kuramına dayanan olabilirlik oran analizi yöntemlerini kullanmıştır. Türkiye örnekleminde cinsiyete yönelik MİF belirlemek için yapılan analizlerde üç yöntemin tutarlı olarak MİF gösterdiğini belirlediği madde bulunmamıştır. Arıkan, Uğurlu, Atar (2016) MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel Yöntemleriyle Gerçekleştirilen MİF ve Yanlılık Çalışması isimli araştırmalarında örneklem büyüklüğü arttıkça SIBTEST yönteminin belirlediği DMF'li madde sayısı arttığını göstermiştir. Ayrıca SIBTEST ve MH yöntemlerine göre DMF içeren maddelere baktığımızda genel olarak iki yöntemde de MİF içeren ortak maddelerin tutarlı olduğu sonucuna ulaşmışlardır.

Yapılan araştırmalarda madde işlev farklılığı belirlemede kullanılan yöntemlerin üstün yanları ve sınırlılıkları tartışılmaktadır. Her bir yöntemin dayandıkları varsayımlar ve matematiksel modeller açısından madde işlev farklılığı belirlemede avantajlı ve dezavantajlı yanları vardır. Bu nedenle bu çalışmada birden çok yöntemin bir arada kullanılıp sonuçların karşılaştırılmasına karar verilmiştir.

Araştırmanın Amacı

Araştırmada, madde işlevinin farklılaşması hakkında istatistiki kanıt bulmak üzere farklı özelliklere sahip olan altı MİF belirleme yöntemine (SIBTEST, Mantel-Haenszel, Rasch Modeli Poly-SIBTEST, Mantel Test, Kısmi Puan Modeli) başvurulmuştur. MİF çalışmalarında birden çok yönteme dayalı inceleme yapılması önerilmektedir (Holland ve Wainer, 1993; Osterlind ve Everson, 2009). Söz konusu yöntemler kullanılarak yapılan analiz sonuçlarının tutarlılığının belirlenmesi alan yazında önem teşkil etmektedir. Türkiye'de MİF belirlemeye yönelik yapılan çalışmalar incelendiğinde, bu çalışmaların daha çok simülatif veri kullanılarak çeşitli koşullar altında MİF belirleme yöntemlerinin nasıl işlediğini gösteren çalışmalar olduğu görülmektedir. Bu araştırma, Türkiye'de psikolojik testlerde, özellikle zekâ testlerinde MİF belirlemeye yönelik yapılan, gerçek veri kullanılan ilk çalışmadır. Bu açıdan bu çalışma daha önce çeşitli koşullar altında denemesi yapılan MİF belirleme yöntemlerinin gerçek veri üzerinde ve psikolojik testlerde test edilmesi açısından önemlidir.

Yukarıda yapılan açıklamalara dayalı olarak bu araştırmanın problemini, Wechsler Çocuklar için Zekâ Ölçeği IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre yanlılık taşıyıp taşımadığını birden fazla yöntemle araştırarak yöntemleri karşılaştırmak oluşturmaktadır. Bu

problem çercevesinde bu araştırmaın amacı, Wechsler Çocuklar için Zekâ Ölçeği IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemlerin (Mantel-Haenszel, SIBTEST ve Rasch MODELİ, Mantel Test, Poly-SIBTEST ve Kısmi Puan Modeli) birbiri ile uyumlu sonuçlar verip vermediğinin karşılaştırılmasıdır.

YÖNTEM

Çalışma Grubu

Bu çalışmada, Wechsler Çocuklar için Zekâ Ölçeği IV Türkiye Uyarlama ve Standardizasyon Çalısması ön uygulama verileri kullanılmıştır. WÇZÖ-IV standardizasyon ön uygulama çalışması doğrultusunda, stardardizasyon çalışmasını yürüten proje ekibinin eğitim verdiği psikologlar tarafından 819 kişiden veri toplanmıştır. Ancak uç değerler nedeniyle 2 kişiye ait veriler analizlere dâhil edilmemiş, analizler 817 kişi üzerinden yürütülmüştür. Testi alan bireylerin sosyo-ekonomik düzeye göre gruplandırılmaları uyarlama çalışmasındaki ölçüte uygun olarak anne eğitim durumları dikkate alınarak yapılmıştır. Okuryazar değil, okuryazar, ilkokul ve ortaokul mezunları düşük; lise mezunları orta; üniversite, yüksek lisans ve doktora mezunları ise yüksek sosyoekonomik düzey olarak gruplandırılmıştır. Testi alanların cinsiyet ve sosyoekonomik düzeye göre dağılımları Tablo 1'de verilmiştir.

Tablo 1. Testi Alan Çocukların Cinsiyet ve Sosyoekonomik Düzeye Göre Dağılımları

Cinsiyet	Frekans	Yüzde
Kadın	443	54.22
Erkek	374	45.78
Toplam	817	100.00
Sosyo-Ekonomik Düzey	Frekans	Yüzde
Yüksek SED	156	19.09
Orta SED	254	31.09
Düşük SED	407	49.82
Toplam	817	100.00

Verilerin Elde Edilmesi

Araştırmada kullanılan veriler, Türkiye'de Wechsler Çocuklar için Zekâ Ölçegi IV (WÇZÖ-IV) Türkiye Norm, Uyarlama ve Standardizasyon Çalışmaları 107K493 ve 109K533 projeleri kapsamında TÜBİTAK destekli olarak Öktem, Gençöz, Erden, Sezgin ve Uluç tarafından (2007-2011 tarihleri arasında) tamamlanan çalışma kapsamında proje ekibinin eğitim verdiği psikologlar tarafından toplanmıştır (Öktem ve diğerleri, 2013). Araştırmada, bu projeler kapsamında toplanan verilerden, Wechsler Çocuklar için Zekâ Ölçeği IV Türkiye Uyarlama ve Standardizasyon Çalısması ön uygulama verileri, proje ekibinin izni alınarak kullanılmıştır.

Veri Toplama Aracı

WÇZÖ-IV David Weschsler'in teorisi temel alınarak geliştirilmiş, bireysel olarak uygulanan bir zekâ testidir. WÇZÖ-IV önceki versiyonlarla karşılaştırıldığında yapısının büyük ölçüde değişime uğradığı görülmektedir. WÇZÖ-R'dan Yapal Zekâ Puanı, Sözel Zeka Puanı ve Tüm Test Zeka Puanı olmak üzere üçtür birleşik puan elde edilmektedir. WÇZÖ-IV için ise Sözel Kavrama Birleşik Puanı (SKBP), Algısal Akıl Yürütme Birleşik Puanı (AAYBP), Çalışma Belleği Birleşik Puanı (ÇBBP), İşlem Hızı Birleşik Puanı (İHBP), Tüm Test Zekâ Puanı (TTZP) olmak üzere beş ayrı birleşik puan elde edilmektedir (Flanagan ve Kaufman, 2004).

Wechsler Çocuklar için Zekâ Ölçeği-IV 10' u asıl, 5'i yedek olmak üzere 15 alttesten oluşmaktadır. Arastırma kapsamında 3 alttest hız testi (Şifre, Simge Arama, Çiz Çıkar) olduğundan analizlere dâhil edilmemiştir. Analizler 6 tanesi çoklu puanlanan, 6 tanesi ikili puanlanan maddelerden oluşan toplam 12 alttest, 315 madde üzerinden yürütülmüştür.

Verilerin Analizi ve Yorumlanması

Öncelikle veriler, cinsiyet ve sosyoekonomik statü göz önüne alınarak düzenlenmiştir. Ardından, Wechsler Çocuklar için Zeka Ölçeği IV Türkçe formunda ikili puanlanan maddeler açısından Rasch modeli, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli ile yapılan analizlerde cinsiyete ve sosyoekonomik düzeye göre MİF gösteren maddelerin tespiti için Rasch modeli kestirimleri yapılmıştır. Rasch modeli doğrultusunda hesaplanan MİF istatistikleri WINSTEP programı ile gerçekleştirilmiştir.

Maddelerin cinsiyete ve sosyoekonomik düzeye göre MİF içerip içermediği MİF kontrast, t ve p değerlerine göre tespit edilmektedir. Bond ve Fox (2001, 2007) maddenin MİF içerdiğinin göstergesi olarak, MİF kontrast±0.5 (MİF Kontrast ≥ |0.5|)'den büyük ve MİF'in istatistiksel olarak anlamlılığının tespiti için t değerinin ±2.0(t≥|2.0|, p<0.05)' den büyük olmak üzere kriterlerin bir arada kullanılmasını önermektedir. Negatif DMF Kontrast değeri maddenin odak gruba kolay geldiğini ve maddenin odak gruba avantaj sağladığını göstermektedir. Bu çalışmada ikili puanlanan maddeler açısından Rasch modeli, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli ile yapılan analizlerde cinsiyete ve sosyoekonomik düzeye göre MİF gösteren maddelerin tespiti için bu kriterler kullanılmıştır.

Son olarak, Wechsler Çocuklar için Zeka Ölçeği IV Türkçe formundaki maddelerin sosyo-ekonomik düzey ve cinsiyete göre işlev farklılığı gösterip göstermediğine karar vermek için analizler WÇZÖ-IV'ün ikili puanlanan maddelerden oluşan alt testleri için Mantel-Haenszel, SIBTEST yöntemleri kullanılarak, WÇZÖ-IV'ün çoklu puanlanan maddelerden oluşan alt testleri için Mantel Test ve Poly-SIBTEST yöntemleri kullanılarak incelenmiştir. Mantel-Haenszel ve Mantel Test'e ilişkin analizler DIFAS programı ile SIBTEST ve Poly-SIBTEST'e ilişkin analizler ise SIBTEST programı ile gerçekleştirilmiştir.

Mantel-Haenszel ve Mantel Test analizlerinde ilgili maddelerin MİF gösterip göstermediğine karar vermek için χ^2_{Mantel} ve $\chi^2_{Mantel\text{-}Haenszel}$ istatistiği kullanılmıştır. Analiz sonucu elde edilen ki-kare istatistiği bir serbestlik derecesinde ki-kare dağılımı göstermektedir (Mantel, 1963; Zwick ve diğerleri,1993; Zwick ve diğerleri, 1997). Bu istatistik için 0.05 anlamlılık düzeyinde kritik değer 3.84, 0.01 anlamlılık düzeyinde kritik değer ise 6.63'tür. Ayrıca Mantel-Haenszel ve Mantel Test analizleri için $\chi^2_{Mantel\text{-}Haenszel}$ ve χ^2_{Mantel} istatistiğine göre MİF gösteren maddelerin hangi alt grup lehine MİF gösterdiğinin belirlenmesi için ikili pualanan maddeler için Mantel-Haenszel-Lojistik Odds Oranı, çoklu puanlanan maddeler için Liu-Agesti- Lojistik Odds Oranı istatistiği kullanılmıştır. Mantel-Haenszel-Lojistik Odds Oranı ve Liu-Agesti- Lojistik Odds Oranının pozitif değerleri o maddenin referans grubun lehine, negatif değerleri ise o maddenin odak grubun lehine çalıştığını gösterir (Penfield, 2013). SIBTEST ve Poly-SIBTEST yöntemi ile MİF gösteren maddeler ve MİF miktarı $|\beta_{UNI}|<0{,}059$ ise maddede ihmal edilebilir düzeyde (A düzeyi); $0{,}059\leq |\beta_{UNI}|<0{,}088$ ise maddede orta düzeyde (B düzeyi); $|\beta_{UNI}|\geq 0{,}088$ ise maddede önemli düzeyde (C düzeyi) ölçütü dikkate alınarak belirlenmiştir. Ayrıca, pozitif β_{UNI} değeri için o maddenin referans grubun lehine, negatif β_{UNI} değeri ise o maddenin odak grubun lehine çalıştığını göstermektedir (Roussos ve Stout, 1996).

BULGULAR

Bu çalışmada verilerin analizi bölümünde belirtilen ölçütlere göre bir maddenin MİF (Madde İşlevsel Farklılığı) içerip içermediğine karar vermek için ikili puanlanan maddeler açısından Rasch

Modeli(MİF Kontrast), Mantel-Haenszel($\chi^2_{\text{Mantel-Haenszel}}$), SIBTEST($\beta_u$) yöntemlerinden, çoklu puanlanan maddeler açısından da Kısmi Puan Modeli(MİF Kontrast), Mantel Test(χ^2_{Mantel}) ve Poly-SIBTEST(β_U) ile yapılan analizlere ilişkin bulgulara yer verilmiştir.

Yukarıdaki ölçütlere göre madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemler (Mantel-Haenszel($\chi^2_{\text{Mantel-Haenszel}}$), SIBTEST($\beta_U$) ve Rasch Modeli(MİF Kontrast), Mantel Test(χ^2_{Mantel}), Poly-SIBTEST (β_U) ve Kısmi Puan Modeli (MİF Kontrast)) açısından her bir alt teste ait cinsiyet açısından MİF içeren maddeler Tablo 2'de, sosyo-ekonomik düzey açısından MİF içeren maddeler ise Tablo 3'de gösterilmiştir.

Tablo 2. İki Koşul İçin Üç Yönteme Göre Her Bir Alt Testteki Cinsiyet Açısından MİF Gösteren Maddeler

Küplerle Desen		
χ^2_{Mantel}	β_u	MİF Kontrast
7	7	-
Benzerlikler		
χ^2_{Mantel}	β_u	MİF Kontrast
5-8-9-21-23	5-8-9	5-23
Sayı Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast
-	-	-
Resim Kavramları		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
12-15	12-15	6-9
Sözcük Dağarcığı		
χ^2_{Mantel}	β_u	MİF Kontrast
20-21-25	17-20-21-25	-
Harf-Rakam Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast
-	6	2-9
Mantık Yürütme Kareleri		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
26	26	-
Kavrama		
χ^2_{Mantel}	β_u	MİF Kontrast
16	16	1
Resim Tamamlama		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
3-6-8-10-15-17-18-19-22-24-32-33	10-15-17-18-19-22-24-32-33	3-6-10-15-17-22-32
Genel Bilgi		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
10-14-17-21	14-17	10-14-17-21
Aritmetik		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
8-10-27	27	8-27

Sözcük Bulma		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
2-4-8-13	-	2-4-8

Madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemler (ikili puanlanan maddeler açısından Mantel-Haenszel, SIBTEST, Rasch MODELİ ve çoklu puanlanan maddeler açısından Mantel Test, Poly-SIBTEST ve Kısmi Puan Modeli) ile yapılan analizlerde, WÇZÖ-IV zeka testinin *ana alt testlerinden; küplerle desen* alt testinde 7. maddede kızlar lehine; *benzerlikler* alt testinde 5., 9., ve 23. maddelerde kızlar lehine; 8. ve 21. maddeler ise erkekler lehine; *resim kavramları* alt testinde 9. ve 12. maddelerde erkekler lehine, 6. ve 15. maddelerde ise kızlar lehine; *sözcük dağarcığı* alt testinde 20 ve 25. maddelerde kızlar lehine, 17 ve 21. maddelerde erkekler lehine; *harf-rakam dizisi* alt testinde 6. maddede erkekler lehine, 2. ve 9. maddelerde kızlar lehine; *mantık yürütme kareleri* alt testinde 26. maddede kızlar lehine; *kavrama* alt testinde 1. ve 16. maddelerde erkekler lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 3., 6., 8., 10., 15., 22., 24. ve 33. maddelerde kızlar lehine, 17., 18., 19., 22. ve 32. maddelerde erkekler lehine; *genel bilgi* alt testinde 10., 14., 17. maddelerde erkekler lehine, 21. maddede kızlar lehine; *aritmetik* alt testi için 8. ve 10. maddeler için kızlar lehine, 27. madde için erkekler lehine; *sözcük bulma* alt testi için 2 ve 8. maddeler için kızlar lehine, 4. ve 13. maddeler için erkeler lehine madde işlev farklılığı bulunmuştur. Ana alt testlerden *sayı dizisi* alt testinde madde işlev farklılığı gösteren maddeye rastlanmamıştır.

Tablo 3. İki Koşul İçin Üç Yönteme Göre Her Bir Alt Testteki Sosyoekonomik Düzey Açısından MİF Gösteren Maddeler

YÜKSEK SED-DÜŞÜK SED			DÜŞÜK SED-ORTA SED			YÜKSEK SED-ORTA SED		
Küplerle Desen			Küplerle Desen			Küplerle Desen		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
14	3-14	-	-	3-4	-	8	4-8	-
Benzerlikler			Benzerlikler			Benzerlikler		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
-	19-23	1-2-4-22	23	-	23	16	-	4
Sayı Dizisi			Sayı Dizisi			Sayı Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
-	13	-	6	6	-	6	6	6
Resim Kavramları			Resim Kavramları			Resim Kavramları		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
16	16	16	6	-	6	13	-	13
Sözcük Dağarcığı			Sözcük Dağarcığı			Sözcük Dağarcığı		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u		χ^2_{Mantel}	β_u	MİF Kontrast
7-8-20-27-32	27-32	1-2-3-4-7-32-33	28	20-28	-	27	27-32	1-2-3-4-7-32
Harf-Rakam Dizisi			Harf-Rakam Dizisi			Harf-Rakam Dizisi		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
-	3-4	1	3	3	9	-	-	-

Mantık Yürütme Kareleri			Mantık Yürütme Kareleri			Mantık Yürütme Kareleri		
$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
-	-	23	22	22	22	22	22	-
Kavrama			Kavrama			Kavrama		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
7-13-16	7-13	3-7-22	2-5-13-22	5-13	22	-	-	3-20
Resim Tamamlama			Resim Tamamlama			Resim Tamamlama		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
11-24-31	11-24-31	5-11-24-26-31-35-36	26	26	26-38	38	38	4-11-36-38
Genel Bilgi			Genel Bilgi			Genel Bilgi		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u	MİF Kontrast
8-20-25	20-25	8-15-20-22-25-27	-	-	6-8-20-28	-	-	13-15
Aritmetik			Aritmetik			Aritmetik		
χ^2_{Mantel}	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast	$\chi^2_{\text{Mantel-Haenszel}}$	β_u	MİF Kontrast
19-23	19-21-23	12-20-23	19-22-28	5-11-19-22-28	7-8-9-11-17-18-19-20-27-28-30-31-32	21-28	21-28	20-21-28
Sözcük Bulma			Sözcük Bulma			Sözcük Bulma		
χ^2_{Mantel}	β_u	MİF Kontrast	χ^2_{Mantel}	β_u		χ^2_{Mantel}	β_u	MİF Kontrast
13	12	8-17-21	12-17	12-17	1-12-17	-	-	-

Madde işlev farklılığını belirlemede kullanılan istatistiksel yöntemler (ikili puanlanan maddeler açısından Mantel-Haenszel, SIBTEST, Rasch MODELİ ve çoklu puanlanan maddeler açısından Mantel Test, Poly-SIBTEST ve Kısmi Puan Modeli) ile yapılan analizlerde yüksek-düşük sosyo-ekonomik düzey karşılaştırıldığında; WÇZÖ-IV zeka testinin *ana alt testlerinden; küplerle desen* alt testinde 14. maddede düşük sed lehine, 3. maddede yüksek sed lehine; *benzerlikler* alt testinde 4., 19. ve 22. maddelerde düşük sed lehine, 1., 2. ve 23. maddelerde yüksek sed lehine, sayı dizisi alt testinde 13. maddede düşük sed lehine; *resim kavramları* alt testinde 16. maddede yüksek sed lehine; *sözcük dağarcığı* alt testinde 1., 2., 3., 4., 7. ve 32. maddelerde yüksek sed lehine, 8., 20., 27. ve 33. maddelerde düşük sed lehine; *harf-rakam dizisi* alt testinde 3. maddede yüksek sed lehine, 1. ve 4. maddelerde düşük sed lehine; *mantık yürütme kareleri* alt testinde 23. maddede yüksek sed lehine; *kavrama* alt testinde 3., 7., 16. ve 22. maddelerde düşük sed lehine, 13. maddede de yüksek sed lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 11., 26., ve 31. maddelerde yüksek sed lehine, 5., 24., 35. ve 36. maddede ise düşük sed lehine; *genel bilgi* alt testinde 8., 20., 22., 25. ve 27. maddelerde düşük sed lehine; 15. madde de ise yüksek sed lehine; *aritmetik* alt testi için 19., 20. ve 23. maddelerde yüksek sed lehine, 21. maddede düşük sed lehine; *sözcük bulma* alt testinde 8., 12., 13. ve 17. maddelerde ise yüksek sed lehine, 21. maddede düşük sed lehine madde işlev farklılığı bulunmuştur.

Orta-düşük sosyo-ekonomik düzey karşılaştırıldığında; WÇZÖ-IV zeka testinin ana alt testlerinden; *benzerlikler* alt testinde 23. maddede düşük sed lehine; *sayı dizisi* alt testinde 6. maddede düşük sed

lehine; *resim kavramları* alt testinde 6. maddede orta sed lehine; *sözcük dağarcığı* alt testinde 20. ve 28. maddelerde düşük sed lehine; *harf ve rakam dizisi* alt testinde 3. ve 9. maddelerde orta sed lehine; *mantık yürütme kareleri* alt testinde 22. maddede düşük sed lehine; *kavrama* alt testinde 5., 13. ve 22. maddelerde orta sed lehine, 2. maddede düşük sed lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 26. maddelerde ise düşük sed lehine, 38. maddede orta sed lehine; *genel bilgi* alt testinde 6., 8., 20. ve 28. maddelerde düşük sed lehine; *aritmetik* alt testi için 5., 9., 11., 17., 18., 20., 22., ve 28., maddelerde düşük sed lehine, 7., 8., 19., 27., 30., 31. ve 32. maddelerde orta sed lehine; *sözcük bulma* alt testinde 12. maddede orta sed, 1. ve 17. maddelerde ise düşük sed lehine madde işlev farklılığı bulunmuştur. Ana alt testlerden *küplerle desen,* yedek alt testlerden ise *genel bilgi* alt testinde madde işlev farklılığı gösteren maddeye rastlanmamıştır.

Yüksek-orta sosyo-ekonomik düzey karşılaştırıldığında ise WÇZÖ-IV zeka testinin ana alt testlerinden; *küplerle desen* alt testinde 8. maddede orta sed lehine; *benzerlikler* alt testinde 4. ve 16. maddelerde yüksek sed lehine; *sayı dizisi* alt testinde 6. maddede yüksek sed lehine; *resim kavramları* alt testinde 13. maddede orta sed lehine; *sözcük dağarcığı* alt testinde 1., 2., 3., 4. ve 27. maddelerde orta sed lehine, 7 ve 32. maddelerde yüksek sed lehine; *mantık yürütme kareleri* alt testinde 22. maddede yüksek sed lehine; *kavrama* alt testinde 3. ve 20. maddelerde orta sed lehine madde işlev farklılığı tespit edilmiştir.

WÇZÖ-IV zeka testinin *yedek alt testlerinden* yine kullanılan yöntemlerin en az ikisine göre; *resim tamamlama* alt testinde 4. maddede yüksek sed lehine, 11., 36. ve 38. maddede orta sed lehine; *aritmetik* alt testi için 21. maddede orta sed lehine, 20. ve 28. maddelerde yüksek sed lehine madde işlev farklılığı bulunmuştur. Ana alt testlerden *harf-rakam dizisi,* alt testinde, yedek alt testlerden ise *genel bilgi, sözcük bulma* alt testinde madde işlev farklılığı gösteren maddeye rastlanmamıştır.

SONUÇLAR ve TARTIŞMA

WÇZÖ-IV'ün maddelerin cinsiyet ve sosyo-ekonomik düzeye göre işlev farklılığı gösterip göstermediğine karar vermek için literatür (Bond ve Fox 2001; Mantel, 1963; Penfield, 2013; Roussos ve Stout, 1996; Zwick ve diğerleri, 1993; Zwick ve diğerleri, 1997) doğrultusundaki ölçütlere göre, her bir alt teste ait cinsiyet ve sosyo-ekonomik düzey açısından üç yönteme göre MİF içeren maddeler gösterilmiştir.

Yöntemler arasında MİF içeren madderin tutarlılığına bakılacak olursa cinsiyet açısından MİF içeren maddelerin tespitinde Mantel-Haenszel, SIBTEST ve Mantel Test, Poly-SIBTEST istatistikleri daha tutarlı sonuçlar verirken sosyoekonomik düzeye göre MİF içeren maddelerin tespitinde Mantel-Haenszel, Rasch MODELİ ve Mantel Test, Kısmi Puan Modelinin daha tutarlı sonuçlar verdiği görülmektedir. Örneklem büyüklükleri ve örneklem eşitsizlikleri aralardaki tutarsızlıkların bir nedeni olarak görülebilir. Bu çalışmanın örnekleminde cinsiyet açısından 443 kız, 374 erkek sosyo-ekonomik düzey değişkeni açısından yüksek sed'den gelen 156, orta sed'den gelen 254, düşük sed'den gelen ise 407 kişi bulunmaktadır. Örnekleme bakıldığında cinsiyet açısından ayrılan grubun örneklem büyüklüğü daha fazla ve gruplar açısından daha dengelidir. Ancak sosyo-ekonomik düzey açısından Türkiye'deki dağılıma da benzer olarak daha dengesiz bir dağılım görülmektedir ve örneklem büyüklüğü daha küçüktür. Bu açıdan bulgular incelendiğinde her üç yöntemin en tutarsız sonuçlar verdiği grubun, örneklem eşitsizliğinin en yüksek olduğu Yüksek SED düşük sed'in karşılaştırıldığı grup olduğu görülmektedir. Bu grupta en çok MİF'li maddeyi Rasch modeli ile tespit edilmiştir. Bu bulgu Taylor ve Lee (2012)'nin karışık madde formatlarında matematik ve okuma testinde cinsiyet yanlılığını Poly-SIBTEST ve Rash modeli ile inceleyen çalışmasının sonuçlarıyla paraleldir. Yine bu grupta SIBTEST, Poly-SIBTEST yöntemi Mantel-Haenszel, Mantel Test yönteminden daha fazla MİF'li madde belirlemiştir.

Awuor'un MİF belirleme tekniklerinin eşit olmayan örneklemlerde gücünü SIBTEST ve Mantel-Haenszel yöntemleriyle test ettiği çalışmasında eşit olmayan örneklem grupları için MH yönteminin I. Tip hatayı SIBTEST yönteminden daha iyi kontrol ettiği sonuçuna ulaşmıştır. Bu sonuca paralel olarak SIBTEST yöntemi ile daha fazla MİF'li maddenin belirlenmesinin I. tip hatadan kaynaklandığı düşünülebilir.

Arıkan, Uğurlu, Atar (2016) MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel Yöntemleriyle Gerçekleştirilen MİF ve Yanlılık Çalışmasında SIBTEST ve MH yöntemlerine göre genel olarak MİF içeren ortak maddelerin tutarlı olduğu sonucuna ulaşmışlardır. Benzer şekilde bu çalışmada da yüksek ve düşük sed'in karşılaştırıldığı grup dışındaki gruplarda yöntemler genel olarak tutarlı sonuçlar vermiştir.

Madde işlev farklılığı testle ölçülen özellik bakımından benzer olup da cinsiyet, sosyoekonomik düzey gibi değişkenler açısından birbirinden farklı alt gruplarda yer alan bireylerin, bir maddeyi doğru cevaplandırma olasılıklarının farklılaşması olarak tanımlanabilir (Hambleton, Swaminathan ve Rogers, 1991). Ancak aynı yetenek düzeyinde olan fakat farklı gruplardan gelen bireylerin, test maddelerine doğru cevap verme olasılıklarının değişmesi maddenin ölçtüğü özelliğin gerçekten bu iki grup arasında farklı olmasından da kaynaklanıyor olabilir. Bu durum madde etkisi olarak adlandırılmaktadır (Clauser, Mzaor, 1998; Mellenberg, 1983; Osterlind, 1983; Shepard, Camilli ve Williams, 1985; Zumbo, Hubley, 1998). Daha önce de bahsedildiği gibi maddelerin, madde yanlılığı içerip içermediğine ilişkin karar verebilmek için maddelerin MİF göstermesi bir gereklilik olmakla birlikte bu kararı vermek için yeterli değildir. MİF olduğu gözlenen maddelerin yanlı olup olmadığının ve yanlı ise bunun nedenlerinin uzmanlar tarafından incelenmesi gereklidir. Bu incelemede uzmanlar tarafından maddenin ölçülmek istenen yapıyla ilişkisiz olarak, bazı alt gruplar için adil olmayan bir avantaj sağlayıp sağlamadığının belirlemesi gerekir (Camilli ve Shepard, 1994; Zumbo, 1999).

Wechsler Çocuklar için Zeka Ölçeği IV (WÇZÖ-IV) 2003 yılında Amerika'da geliştirilip kullanıma sunulmuş, 2011 yılında ise Türkiye'de uyarlama çalışması tamamlanarak kullanılmaya başlanmıştır. Hambleton ve arkadaşlarının editörlüğünü yaptığı Adapting Educational and Psychological Tests for Cross-Cultural Assessment isimli kitabında uyarlanmış testlere ilişkin hataların ve geçerlik sorunlarının kaynağını kültürel ve dilsel farklılıklar, dizayn, metot gibi teknik konular ve sonuçların yorumlanması olmak üzere üç başlık altında toplamıştır. Kültürel ve dilsel farklılıklar yapıların eşdeğer olmaması, testin yönergesinin verilmesi ve testin uygulanışı ile ilgili sorunlar, madde formatları ve testi alanların hızlarının etkisi faktörlerinden etkilenmektedir. Dizayn, metot gibi teknik konular ise çevirmenlerin seçimi ve eğitimi, çeviri süreci, veri toplama süreci faktörleriyle ilişkilidir. Sonuçların yorumlanması ise eğitim programının benzerliği, öğrencinin motivasyonu, sosyo-politik özellikler gibi faktörlerden etkilenmektedir (Hambleton, 2005). Görüldüğü üzere bir testin uyarlama sürecinde testin geçerliğini etkileyen bir çok faktör vardır. Testin uyarlandıktan sonra yanlılık içermesi de testin geçerliğini tehdit eden sorunlardan biridir. Bir teste ilişkin yanlılık türleri yapı yanlılığı, metot yanlılığı ve madde yanlılığı (madde işlev farklılığı) olarak ayrılabilir. Bu çalışmada ele alınan madde yanlılığının nedenleri ise zayıf çeviriler, muğlak birden çok anlama gelen ifadeler, maddenin ilişkili olduğu istenmeyen faktörler (örnek olarak maddenin ölçülmek istenen yapı dışında başka bir yapıyı ölçmesi) olarak sıralanabilir (Vijver, Poortinga, 2005). Bu araştırmada, belirlenen farklı yöntemlere göre MİF gösteren maddelerin saptanıp bu yöntemler arasındaki uyum düzeyleri ortaya konulmuştur. Bundan sonraki çalışmalar adına MİF bulunan maddelerin uzmanlarca incelenerek yanlılık içerip içermediği belirlenmeli, cinsiyet ve sosyo-ekonomik düzeye göre MİF içeren maddeler için yanlılık içerdiği tespit edilir ise bu yanlılığın kaynağı ve nedenlerine ilişkin yanlılık gösteren maddeler, yukarıda belirtilen tüm faktörler göz önüne alınarak yorumlanmalı uzmanlar tarafından ayrıntılı olarak incelenmelidir.

KAYNAKÇA

Ackerman, T. A., & Evans, J. A. (1992). *An investigation of the relationship between reliability, power, and the Type I error rate of the Mantel-Haenszel and simultaneous item bias detection procedures.* Annual Meeting of the National Council on Measurement in Education, San Francisco.

Adams, R. J., & Rowe, K. J. (1988). Test bias. In J. P. Keeves (Ed), *Educational teearch, methodology, andmeasurement: An international handbook* (p. 398-403). Oxford: Pergamon Pres.

Arıkan, Ç., Uğurlu, S. ve Atar, B. (2016). MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-Haenszel yöntemleriyle gerçekleştirilen DMF ve yanlılık çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(1), 34-52 .

Atalay, K., Gök, B., Kelecioğlu, H. ve Arsan, N. (2012). Değişen madde fonksiyonunun belirlenmesinde kullanılan farklı yöntemlerin karşılaştırılması: Bir simülasyon çalışması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *43*, 270- 281.

Awuor, R. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-Based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures* (Unpublished doctoral dissertation), Virginia Polytechnic Institute and State University Blacksburg, Virginia, US.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* New Jersey: Lawrence Erlbaum Associates.

Camili, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. London: Sage.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333-353.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th Ed.). New York: Harper Collins Publishers.

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

Gierl, M. J., Jodoin, M., & Ackerman, T. (2000). *Performance of Mantel-Haenszel, simultaneous item bias test and Logistic Regression when the proportion of DIF items is large.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana, USA.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. CA: Sage.

Henderson, D. L. (1999). *Investigation of differential item functioning in exit Examinations across item format and subject area* (Unpublished doctoral dissertation). University of Alberta, Edmonton, Alberta, Canada.

Holland P. W., & Wainer, H. (1993). *Differential Item Functioning*. New Jersey: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel- Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horst, P. (1966). *Psychological measurement and prediction.* Belmont: Wadsworth Pub. Co. Institute for Education Research.

Kamata, A., & Vaughn, K. B. (2004). An introduction to Differential Item Functioning analysis. *Learning Disabilities: A Contemporary Journal, 2*(2), 49-69.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

Mantel, N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute, 22*, 719- 748.

Mellenberg, G. J. (1983). Conditional item bias methods. In S. H. Irvine & W. J. Barry (Eds.), *Human assesment and cultural factors* (p. 293–302). New York: Plenum Pres.

Mellor, T. L. (1995). *A comparison of four differantial item functioning methods for polytomously scored items* (Unpublished doctor dissertation). The university of Texas, Austin.

Murphy, K., & Davidshofer, C. (1994). *Psychological testing: Principles and applications* (3th Ed.). Englewood Cliffs, NJ: Prentice Hall.

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform bias. *Applied Psychological Measurement, 20*(3), 257-274.

Osterlind, S. J. (1983). *Test item bias.* California: Sage.

Osterlind S. J., & Everson H. T. (2009). *Differential Item Functioning* (2nd Ed.). California: Sage.

Öktem, F., Gençöz, T., Erden, G., Sezgin, N. ve Uluç, S. (2013). *Wechsler çocuklar için zeka ölçeği-IV (WÇZÖ-IV) uygulama ve puanlama el kitabı Türkçe sürümü.* Türk Psikologlar Derneği Yayınları, Ankara.

Penfield, R. (2013). *DIFAS 5.0 Differential item functioning analysis system: User's manual*. http://soe.uncg.edu/wp-content/uploads/2015/12/DIFASManual_V5.pdf

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197–207.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish

Reynolds. C. R., Livingston R. B., & Willson, V. (2006). *Measurement and assessment in education.* Boston: Pearson Education. Inc.

Rodney. G. L., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology. 75*, 164- 174.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116.

Roussos, L. A. (1992). *Hierarchical agglomerative clustering computer program user's manual.* University of Illinois at Urbana-Champaign.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215-230.

Shealy, R., & Stout, W. (1993). An item response theory model for test bias and differential test functioning. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (p. 197–240). Hillsdale, NJ: Earlbaum.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77–105.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*(3), 246-280.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.

Tittle, C. K. (1988). Test bias. In J. Keeves (Ed.), Educational teearch, methodology, and measurement: An international handbook (p. 392- 398). Pergamon Press: UK.

Ulutaş, S. (2012). PISA 2006 fen okuryazarlığı testindeki maddelerin yanlılık bakımından araştırılması (Yüksek Lisans Tezi). Ankara Üniversitesi, Eğitim Bilimleri Enstitüsü.

Uttaro, T., & Millsap, R. E. (1994). Factors in uencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

Vijver, F. ve Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (p. 39-63). Mahwah, NJ: Lawrence Erlbaum.

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.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zwick, R., Donoghue, J. R., Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233- 251.

EXTENDED ABSTRACT

Introduction

Bias in statistical terms is a systematic and constant error of measurement as opposed to chance error or a systematic under or over estimation of a population parameter by a statistic based on samples which are drawn from the population (Camilli ve Shepard, 1994; Zumbo 1999). It is unacceptable for educational and psychological tests to contain biased items. If some items provide an advantage to a specific subgroup of examinees that increases their ability to respond correctly, that is, if some items exhibit systematic bias, it is impossible to say that the test was equally fair for all examinees. A typical approach to investigate bias at the item level is called differential item functioning (DIF)

analysis, defined as a difference in the measurement properties of an item for demographic subgroups (Camilli, Shepard, 1994).

Differential item functioning identifies differences in the probability of answering an item correctly accordingly for identifiable subgroups, at every ability level of psychological structure that is aimed to be measured with an item (Embretson & Reise, 2000; Lord, 1980). Statistical procedures that are currently used by test publishers to identify items that function differentially tend to focus on such subgroups as gender, race, language, or socioeconomic status groups. DIF analyses are useful for flagging items that may need to be eliminated or, at least, submitted to additional review.

In the beginning of 1900's, it was recognized that some items that were used to measure IQ were also measuring the effects of cultural training instead of mental capacity. Group differences in IQ tests have been the major research area for many researchers.

The Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Wechsler, 2003) is an individually administered IQ test for children. The Wechsler scales are among the most adapted tests in the world. In developing or adapting a standardized test that is fair for all test takers requires the removal or the revision of potentially biased items. When practitioners are developing or adapting instruments, they should conduct DIF analysis to investigate possible bias at the item level.

The purpose of this study is to investigate potential gender and socio-economic status bias in the WISC-4 by using several differential item functioning detection techniques. In addition, this study will show similarities and differences in practice by comparing three DIF-detection techniques: the Rasch item response theory (IRT) estimations, SIBTEST (Shealy, Stout, 1993), and the Mantel-Haenszel techniques (Holland, Thayer, 1988).

Method

In this study, Wechsler Intelligence Scale for Children: Fourth Edition Turkish standardization pilot test data were used. In accordance with the Wechsler Intelligence Scale for Children: Fourth Edition standardization study, pilot data have been collected from 817 children by psychologists. Table 1 shows descriptive statistics broken down by gender and socio-economic status.

Table 1. Descriptive Statistics for Gender and Socio-Economic Status

Gender	Frequency	Percentage
Female	443	54.22
Male	374	45.78
Total	817	100
Socio-Economic Status	Frequency	Percentage
High	156	19.09
Moderate	254	31.09
Low	407	49.82
Total	817	100

The WISC-4 is an individually administered intelligence test, based on the theory of David Wechsler. Different from previous versions of the Wechsler tests, the structure of WISC-IV is observed to have undergone a significant change, approximately 45% of the entire test has been revised (Flanagan and Kaufman, 2009). WISC-R gives three types of combined scores: a Performance Intelligence Score, a Verbal Intelligence Score, and a Full Test Intelligence Score. However, the WISC-IV also provides additional indexes: namely a Verbal Comprehension Index (VCI), a Perceptual Reasoning Index (PRI), a Working Memory Index (WMI), a Processing Speed Index (PSI), and a Full Scale IQ (FSIQ).
In accordance with the purpose of the study, 315 items were used both in polytomously scored subtests such as *Block Design, Similarities, Digit Span, Vocabulary, Letter-Number Sequencing,*

Comprehension, and dichotomously scored subtest such as *Picture Concepts, Matrix Reasoning, Picture Completion, Information, Arithmetic*, and *Word Reasoning.* Three subtests were excluded from this research because they involved a speeded format.

First, data were prepared for analyses. Then Rasch and Partial Credit Model analysis were performed to detect differential item functioning. DIF values have been estimated using the IRT estimation software Winsteps. Bond and Fox (2001, 2007) suggest t value ± 2.0 ($t \geq |2.0|$, $p<0.05$) and DIF Contrast ± 0.5 ((DIF Contrast ≥ |0.5|) asDIF indicators based on the studied groups. In this research, these criterions were used for evaluating the items if they have DIF or not.
In addition to Rasch DIF analysis, DIF analyses were performed using two different techniques: SIBTEST-Poly-SIBTEST and Mantel-Haenszel Test-Mantel Test. SIBTEST-Poly-SIBTEST analyses have been computed using the SIBTEST software. To decide DIF according to SIBTEST results, β_{UNI} can be interpreted as the magnitude of DIF for each item. Positive values of β_{UNI} indicate DIF favoring the reference group, and negative values indicate DIF favoring the focal group. Roussos and Stout (1996) proposed guidelines for interpreting DIF by combining the SIBTEST statistical results with values for the β_{UNI} parameter estimate to classify DIF on a single item: (a) negligible DIF ($\beta_{UNI} < 0.059$ and H_0: $B = 0$ is rejected), (b) moderate DIF ($0.059 \leq \beta_{UNI} < 0.088$ and H_0: $B = 0$ is rejected), (c) large DIF, ($\beta_{UNI} > 0.088$ and H_0: $B = 0$ is rejected).
Mantel-Haenszel and Mantelanalyses had been computed using the DIFAS software. To decide DIF according to Mantel-Haenszel Test and Mantel Test results, critical values of this statistic are 3.84 for a Type I error rate of 0.05 and 6.63 for a Type I error rate of 0.01. The Mantel chi-square statistic (Mantel, 1963, Zwick, Donoghue & Grima, 1993; Zwick, Thayer & Mazzeo, 1997) is distributed as chi-square with one degree of freedom. To indicate DIF favoring group, Liu-Agresti cummulative common log-odds ratio (Liu & Agresti, 1996; Penfield & Algina, 2003) and *Mantel-Haenszel Common Log-Odds Ratio* (Camilli & Shepard, 1994; Mantel & Haenszel, 1959) statistics were used. This statistic is asymptotically normally distributed. Positive values indicate DIF in favor of the reference group, and negative values indicate DIF in favor of the focal groups.

Results and Discussion

In this reseach, for investigating potential gender and SES bias in the WISC-4, three DIF detection techniques were used. The items that are detected as DIF items based on gender and SES are summarized on Tables 2 and 3. In addition, this study has shown similarities and differences in practice by comparing three DIF-detection techniques: The Rasch item response theory (IRT) estimations, SIBTEST (Shealy, Stout, 1993), and the Mantel-Haenszel techniques (Holland, Thayer, 1988). In conclusion, as it is shown in Table 2 and Table 3, potential gender and SES biased items were identified according to three DIF-detection techniques. Even though items detected as DIF items with three techniquess were quite similar, there are also differences based on these three techniques. There are many reasons for differences based on techniques in the detection of DIF items. Also in order to detect and/or prevent bias, we need to recognize factors that can induce bias. It is important to understand from a substantive, cognitive perspective why these gender and SES differences and also differences based on techniques in the detection of DIF items are occurring.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);128-145

An Investigation of Group Invariance in Test Equating According to Gender

Test Eşitlemede Grup Değişmezliğinin Cinsiyete Göre İncelenmesi

Hatice İNAL * Çiğdem AKIN ARIKAN **

Abstract
The aim of this study is to investigate the group invariance condition according to Tucker and Levine observed score equating among linear equating methods. In the study, the 4th and 6th booklets of the PISA 2012 Mathematics subtest were used. Booklets were equated according to group and gender sub-variables, and then group invariance of each condition and WMSE values were calculated. Within this scope, REMSD and RMSD (x) group invariance indexes were employed. The results of the study indicated that, when WMSE values, obtained according to equating methods, were compared, Tucker observed score equating method with regard to whole-group and gender sub-groups produced the lowest error. When RMSD and REMSD values obtained according to gender sub-groups were examined by linear equating methods, it was found that group invariance value is smaller than criterion value for Tucker equating method, while it was greater than criterion value for Levine equating method. Eventually, group invariance condition was met for Tucker observed score equating, but not for Levine observed score equating.

Keywords: Equating, group invariance, Tucker linear equating, Levine linear equating

Öz
Bu çalışmanın amacı doğrusal eşitleme yöntemlerinden Tucker ve Levine gözlenen puan eşitleme yöntemlerine göre eşitlemenin grup değişmezliği koşulunun incelenmesidir. Bu çalışmada PISA 2012 matematik alt testine ait 4. ve 6. kitapçıklardan elde edilen test puanları kullanılmıştır. Kitapçıklardan elde edilen puanlar tüm grup ve cinsiyet alt değişkenine göre eşitlenmiştir. Her bir koşula ait grup değişmezliği ve WMSE değerleri hesaplanmıştır. Bu bağlamda grup değişmezliği indekslerinden REMSD ve RMSD (x) kullanılmıştır. Araştırma sonucunda eşitleme yöntemlerine göre elde edilen WMSE değerleri karşılaştırıldığında hem tüm grup hem cinsiyet alt grubuna göre en az hata veren yöntemin Tucker gözlenen puan eşitleme olduğu görülmüştür. Doğrusal eşitleme yöntemleriyle cinsiyet alt grubuna göre elde edilen RMSD ve REMSD değerleri incelendiğinde, Tucker eşitleme yöntemi için kriter değerden küçük iken, Levine eşitleme yöntemi için kriter değerden daha yüksek çıkmıştır. Böylece grup değişmezliği koşulunun Tucker eşitleme yönteminde sağlanırken, Levine eşitleme yönteminde sağlanmadığı görülmüştür.

Anahtar Kelimeler: Eşitleme, grup değişmezliği, Tucker eşitleme, Levine eşitleme

INTRODUCTION

PISA (Programme for International Student Assessment), that enables countries to compare their educational indicators, was administered by OECD in every three years since 2000. PISA application assesses to the extent which students at the age group of 15 are equipped with the basic mathematics, science and reading knowledge and skills in order to help them be a part of the modern society. PISA application aims to determine the extent students' ability to utilize knowledge and skills to use them in real life, understand the new situations, resolve problems, make guesses about what they are unfamiliar with and make judgments. In PISA application, students are required to take the all test item sets that consist of science, mathematics and reading skills. The items sets are incorporated in 13

* Araş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: hinal@hacettepe.edu.tr
** Araş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Ankara-Türkiye, e-posta: akincgdm@gmail.com

DOI: *10.21031/epod.301326*
Geliş Tarihi: 23.11.2016
Kabul Tarihi: 06.03.2017

booklets and there are some common items to link all the booklets (OECD, 2014). Therefore, it is necessary to equate the scores in order to compare these scores obtained from different booklets.

Equating can be described as the statistical process, which regulates the differences between the tests, forms with the same content and difficulty level and enables the scores obtained from these form to be used interchangeably (Kolen, 1988). The aim of test equating is to make sure that the difficulty of the test form does not create any advantage or disadvantage to the test taker. There are some conditions that must be met in order to equate the test forms. These conditions include equality, symmetry, group invariance and unidimensionality (Hambleton & Swaminathan, 1985). Among these conditions, group invariance means equating function is independent from the sub-groups so that sub-groups do not affect the equating (Kolen, 2004). For example, when two forms of a test are equated, it is possible to obtain the same equated scores for the female and males only when the group invariance condition is met. When group invariance is not ensured, students with different gender and same skills can obtain different equated scores and thus students have advantage or disadvantage because of their genders. In other words, it is fair to say that group invariance is related to equality and objectivity in assessment and evaluation (Dorans, 2004, 2008). In literature, different group invariance criteria have been developed to assess the accuracy as well as fairness of equated scores. These criteria are based on controlling the correspondence of equation principles (Petersen, Kolen & Hoover, 1989).

Test equating is divided into two groups as traditional and Item Response Theory approaches. Traditional equating methods include mean equating, linear equating and equipercentile equating methods (Kolen & Brennan, 2004). Mean equating is based on the assumption that test forms differ with respect to difficulty levels and this difference is fixed across whole scale. For example, in mean equating how much did responders in the upper group found X form easier than Y form will be the same for the individuals in the lower group (Kolen & Brennan, 2014). The equation of mean equating is as follows:

$$m_y(x) = y = x - \mu(X) + \mu(Y) \quad (1)$$

If reference and score distribution of the new form are not equal, equipercentile equating method is used. It is accepted that in the score distribution of X and Y forms, the scores that correspond to the same percentile rank are equal. Equipercentile equating consists of two steps. First, cumulative frequencies of two forms are transferred to a table and cumulative frequency table is drawn. Second the scores that correspond to the same percentile rank are equated. With the scores that are obtained via equipercentile equating method, score distribution of the new form and reference form becomes similar (Livingston, 2004; Kolen, 1988).

When features of two test forms are the same except from means and standard deviations, linear equating is used (Crocker & Algina, 1986; Kolen & Brennan, 2014). In other words, the scores that correspond to the same standard scores (Z scores) are accepted as equal. If the standard deviations of test forms are equal, linear and mean equating will yield the same results. If raw scores and equated scores are given in the same graph, their linear relationship can be illustrated. Linear equating equation is presented in equation 2.

$$\frac{y-\bar{Y}}{s_y} = \frac{x-\bar{X}}{s_X} \quad (2)$$

In linear equating, if the groups, which take the forms differ in terms of their skills, anchor items are used. Different linear equating methods have been developed to equate the forms, which have common items (Livingston, 2004).

Linear Equating Methods for the Non -Equivalent Groups

Non-Equivalent groups Anchor Test-NEAT, the common items pattern, is administered when it is not possible to administer the test form more than once due to test reliability in non-equivalent groups. In NEAT pattern, both forms incorporate some common items and these forms are administered on the

non-equivalent groups. Equating relationship between the test forms is established via common items. Common items are classified as internal and external. If the score obtained from the common items is added to the test score of the test taker, it is called internal anchor, if not, it is called as external anchor. In linear equating for the NEAT pattern, equating relationship prediction is made over a single group by combining non-equivalent Group 1 and Group 2. Braun & Holland (1982) & Angoff (1971) named this group as synthetic. Group 1 and Group 2 classified as synthetic are weighted with w1 and w2. Weighting has two rules. The first of these rules is that the sums of two weights are completely equal (w1+w2=1) and the second one is that each weight equals to zero or is bigger than zero (w1, w2 ≥ 0). Even tough w1=w2=0,5 where two weights are equal are used in general, synthetic is used in (w1=1, w2=0) when group is only defined as new (Topczewski, Cui, Woodruff, Chen & Fang, 2013; Kolen & Breannan, 2014). In this study, the case in which both weights are equal was used.

Equation for linear equating in non-equivalent groups on common items pattern $l_{y_s}(x)$ is the equation used for equating the X observed scores with Y observed scores and s stands for the synthetic group):

$$l_{y_s}(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}[x - (\mu_s(X)] + \mu_s(Y) \quad (3)$$

$\mu_s(X)$ stands for the mean score of the new form obtained from the synthetic group, $\mu_s(Y)$ stands for the mean score of the reference form obtained from the synthetic group; $\sigma_s(Y)$ stands for the standard deviation of the reference form obtained from the synthetic group, $\sigma_s(X)$ stands for the standard deviation of the new form obtained from the synthetic group.

Four parameters of synthetic population in Equation 3, are indicated by the following Equations No. 4, 5, 6 and 7 for Group 1 and Group 2.

$$\mu_s(X) = w_1\mu_1(X) + w_2\mu_2(X) \quad (4)$$
$$\mu_s(Y) = w_1\mu_1(Y) + w_2\mu_2(Y) \quad (5)$$
$$\sigma_s^2(X) = w_1\sigma_1^2(X) + w_2\sigma_2^2(X) + w_1w_2[\mu_1(X) - \mu_2(X)]^2 \quad (6)$$
$$\sigma_s^2(Y) = w_1\sigma_1^2(Y) + w_2\sigma_2^2(Y) + w_1w_2[\mu_1(Y) - \mu_2(Y)]^2 \quad (7)$$

In non-equivalent groups, common items pattern $\mu_s(X), \mu_s(Y), \sigma_s^2(X)$ and $\sigma_s^2(Y)$ cannot be calculated directly since Group 1 does not take X form and Group 2 does not take Y form. Therefore, some assumptions are required according to the equating methods used (Kolen & Brennan, 2004).

Linear equating methods that are used in non-equivalent groups common items pattern can be listed as Levine observed score equating, Levine true scores equating, chained linear equating, Braun-Holland Linear equating (Kolen & Brennan, 2014). Since a group can take only one form in non-equivalent groups common items pattern, linear equating also requires powerful statistical assumptions (Chen, Cui, Zhu & Gao, 2010). In this study, since Tucker and Levine observed score equating methods were used, only information about them was mentioned.

Tucker observed score equating

Tucker method was defined by Gulliksen in 1950 (Kolen & Brennan, 2014). The assumptions required for Tucker observed score equating method are related to regression and conditional variance. The first assumption requires the regression on the common item scores of total scores within both samples are equal. Conditional variance assumption requires variances of the total scores conditions are equal for both samples (Chen et al, 2010; Kolen & Brennan, 2014).

Levine observed scored equating

Levine originally developed the method in 1955 without considering the concept of a synthetic population. After improvements, this method became more general than Levine's (1955).

There are three assumptions of Levine observed score equating.

I. X, Y and common items measure the same characteristics and real scores of X, Y and common items are interlinked within both groups.

II. The regression of X and Y forms on common items are linear and equal within both groups.

III. Error variance of X and Y forms is equal within both groups (von Davier & Kong, 2003; Kolen & Brennan, 2014).

Group invariance is one of important condition to provide test score interchangeability. If group invariance wasn't met, it can be said that the equating couldn't be performed satisfactorily. Multiple studies that investigate the group invariance condition according to different subgroups s available (Dorans, 2004; Yang, 2004; von Davier & Han, 2004; Yin, Brennan & Kolen, 2004; von Davier & Wilson, 2008; Yang & Gao, 2008, Yi, Harris & Gao, 2008; Dorans, Liu & Hammond, 2008). However, although there are plenty of studies related to equating in Turkey (Kelecioğlu,1994; Şahhüseyinoğlu, 2005; Bozdağ & Kan, 2010; Kan, 2011; Kilmen, 2010; Gök, 2012; Öztürk, 2010; Kahraman, 2012; Kelecioğlu & Öztürk Gübeş, 2013; Mutluer, 2013; Demir & Güler, 2014; Atalay Kabasakal, 2014; İnci, 2014; Uysal, 2014), there is no more research which investigated group invariance of equating results. (Öztürk-Gübeş, N. & Kelecioğlu;2017).

The aim of this study is to equate test scores using Tucker and Levine observed score equating methods among linear equating methods according to non-equivalent groups common items pattern and to investigate whether or not group invariance condition of equating methods is met with respect to gender sub-groups. Additionally, in order to assess score equating, this study addressed how group invariance was applied by using real data.

Sub-problems

The purpose of the study is to investigate group invariance of the equated scores obtained from Tucker and Levine observed score equating method with respect to gender. For this purpose, these research questions were examined

1. How the results of Tucker and Levine are observed score equations for total score?
2. How the results of Tucker and Levine are observed score equating with regard to gender?
3. How the results of group invariance according to Tucker and Levine are observed score equating methods?
4. Which is the better option from Tucker and Levine equating methods to equate the test forms?

METHOD

This study aims to equate two booklets administered in Turkey in PISA 2012 (4^{th} and 6^{th} booklets) and assess the equating results. Therefore, this study can be considered as descriptive since the existing method and techniques were assessed via real data.

Population and Sampling

A total of 510 thousand students at the age of 15 participated in PISA application as the representatives of 28 million students from 65 countries in 2012. 4848 students from Turkey participated in PISA in 2012. The sample of the study consists of 741 students, who took 4^{th} and 6^{th} booklet of PISA in Turkish Descriptive statistics of the sample are presented in Table 1.

Table 1. Descriptive Statistics Regarding Gender

Descriptive statistics						
Booklet	Gender	N	Mean	Standard deviation	Skewness	Kurtosis
Booklet 4	Female	182	13,302	6,873	,715	-,036
	Male	197	13,944	8,047	,555	-,719
Booklet 6	Female	178	12,601	7,088	,648	-,284
	Male	184	13,647	7,728	,768	-,082

Data Collection Tools

For data analysis, the data set of the mathematical literacy items by the Turkish students who participated into PISA 2012 application was used. There were 13 booklets in PISA 2012 application. The 4th and 6th booklets were used in this study. 4th booklet included 37, 6th booklet included 36 items. Since traditional equating methods are used in the present study, the most difficult item was excluded from the 4th booklet and the number of the items was equated. The data used in this study were downloaded from official website of OECD (http://pisa2012.acer.edu.au/). Later, correct answers, wrong answers and missing data were coded as 1, 0 and 0, respectively and all partially correct and correct answers to a couple of partially scored items were coded as 1 and the data to be analyzed was made ready.

Data Analysis

Data analysis was conducted at four steps. At the first step, it was examined whether or not the booklets met the equating conditions, at the second one the equated scores were obtained by using different equating methods, at the third one group invariance indexes were calculated in order to see how equating function obtained by each equating method differed across groups and at the final step error in each equating method was calculated.

I. Step: At the first step of data analysis, it was examined whether or not equating conditions are met.

To this end, primarily it was tested if the data was unidimensional. Tetrachoric correlation based principal components factor analysis, is used in order to determine the unidimensionality. This analysis was conducted with Factor 9.3 (2014) program developed by Lorenzo-Seva.

Table 2. Results of the factor analysis

	Booklet 4		Booklet 6	
Component	Eigenvalue	P.E.V (%)	Eigenvalue	P.E.V (%)
1	8.884	0.246	8.597	0.238
2	1.764	0.049	1.565	0.434

P.E.V (%): Proportion of explanation variance

The results of the factor analysis presented in Table 2 demonstrate that there is more than 4 times decline between the 1st factor and the 4th factor and the explanation variance of the second factor was quite low. Therefore booklets have a single general factor, which implies the tests meet the unidimensionality assumption.

Ratio test was administered in order to determine whether or not there was a significant relationship between the average difficulties of the forms (Baykul, 1996). The results of the test to compare the average difficulty of the booklets are presented in Table 3.

Table 3. Comparison of the average difficulty of the booklets

Booklets	$\bar{p}$	T	p
4	0.368	0.638	0.738
6	0.364		

*p<0.05

When Table 3 is examined, there is no statistically signifcant difference between difficulty levels of booklets (p>.05). In this case, the equality of average difficulty of the booklets to be equated, which is another condition for equation, is ensured.

KR-20 reliability coefficient was calculated in order detect if the booklets to be equated are equally reliable. Fischer's Z statistics was carried out in order to detect if there was a difference between two reliability coefficients (Akhun, 1984). The findings regarding the differences in reliability coefficients are presented in Table 4.

Table 4. Comparison of the reliability of booklets

Booklets	KR-20	Z_r	Z	p
4	0.905	1.499	0.367	0.643
6	0.900	1.472		

When Table 4 is examined, it is seen that there is no significant difference between the reliability of the booklets at .05 alpha level/%95 confidence interval (p>.05). This demonstrates that the booklets meet the equal reliability condition.

T test and Levine test were used to test difference between the mean scores and variances of the booklets, respectively. The findings regarding the analyses are presented in Table 5.

Table 5. Comparison of the means and variances of the booklets

		t test			Levene's test		
Booklets	N	$\bar{X}$	t	p	S^2	F	p
4	379	13.635	0.917	0.359	56.300	0.665	0.415
6	362	13.132			55.184		

When Table 5 is examined, it is seen that there is not a significant difference between the means and variances of the booklets at .05 level.

At the end of the analyses regarding the necessary conditions for equating, it was seen that the tests are one-dimensional, are equal in reliability, variances and average difficulty.

II. Step: At the second step of the data analysis, equated scores were obtained by using Levine and Tucker equating methods. Tucker and Levine observed score equating was performed in Microsoft Excel program.

III. Step: At the third step of the data analysis, group invariance indexes were calculated in order to assess whether equated scores obtained via each equating method differed in female and male subgroups. In this study, RMSD(x) and REMSD indexes developed by Dorans and Holland (2000) were employed in order to determine group invariance.

RMSD(x) (Root Mean Square Difference): The value found by RMSD(x) denotes the distance between the subgroup equating functions and total equating function at a x score level. In literature, studies indicating that RMSD(x) can be adapted to other equating method and patterns are available (von Davier, Holland & Thayer, 2004; von Davier & Wilson, 2008). These studies indicate that RMSD(x) can be reported in the form of other equating methods and patterns by eliminating the denominator of the equation in an unstandardized way.

x: Determined score level of the test form

j: Subgroup level

$e_{pj}(x) - e_p(x)$**:** The difference between the equated score calculated based on the equating function of the subgroup j at an x score level with the equated score calculated based on the total equating function

w_j**:** The weight that is determined with the help of the ratio of the test-takers with the subgroups for each subgroup

σ_{YP}**:** Standard deviations of the scores in Q group (Q stands for the one and only group that is examined in single-group or random groups pattern) are defined with the following equation

$$RMSD(X) = \frac{\sqrt{w_j\left[e_{pj}(x) - e_p(x)\right]^2}}{\sigma_{YP}} \quad (8)$$

and with the help with this equation, it is possible to determine group invariance in case of single-group or equivalent groups equating pattern and linear equating function (Dorans & Holland, 2000).

$$RMSD(X) = \sqrt{w_j\left[e_{pj}(x) - e_p(x)\right]^2} \quad (9)$$

Dorans and Holland (2000) described the score level independent state of RMSD(x) as REMSD (Root Expected Mean Square Difference).

$$REMSD(X) = \frac{\sqrt{w_j E_p\left[e_{pj}(x) - e_p(x)\right]^2}}{\sigma_{YP}} \quad (10)$$

In this equation, E_p stands for the mean score of the distribution found with the help of the differences between the equated scores. A group invariance study yields one REMSD. In literature, some studies stating that REMSD can be adapted to other equating method and patterns are available (von Davier, Holland & Thayer, 2004; von Davier & Wilson, 2008). These studies indicate that RMSD (x) can be reported in the form of other equating methods and patterns by eliminating the denominator of the equation in an unstandardized way.

In assessing the group invariance in equating, DTM criterion, which is taken as the half of the raw score unit and recommended by Dorans, Holland, Thayer & Tateneni (2003) and Dorans (2004) is utilized. It is not a certainly set rule to assess the group invariance based on DTM scope. In this study interpretations were made by considering that the difference smaller than 0.50 between equated score of the whole group and the equated score of a sub-group(s) is negligible and difference bigger than 0.50 is significant (Kolen & Brennan, 2014).

IV. Step: At the final step of the data analysis, error of each equating method was calculated. In this study, weighted mean squares error (WMSE) was used in order to assess equating error.

WMSE (Weighted Mean Squares Error): It is used in order determine which method is the most suitable in line with the error of the scores equated according to different equating methods. Weighted mean squares error (WMSE) is calculated by comparing the equated scores corresponding to each raw score at the same skill level (Skaggs & Lissitz, 1986). Skaggs and Lissitz (1988) reported that WMSE

index is quite similar to the total error indexes available in other equating studies. The equation for the calculation of WMSE coefficient is given below:

$$WMSE = \frac{\sum_{i=1}^{k-1} fi(X_E - X_{Crit})^2}{\sum_{i=1}^{k} f_i S^2 y} \quad (11)$$

k : The number of the items in Y test.
S^2y**:** Variance of the raw scores in Y test.
$\mathbf{X}_{crit:}$ i. raw score in Y test.
$\mathbf{X}_E$: the score obtained via equating methods and that correspond to i. raw score in X test.
fi: i. raw score frequency in Y test

FINDINGS

The equated scores of PISA 2012 Mathematics sub-test obtained for Tucker and Levine observed score equating methods with respect to gender and the raw scores are presented in Table 6. The graphs regarding the raw scores obtained for both methods and equated scores are given in the appendix.

Table 6. Raw scores and the scores that correspond to these scores that are obtained via Tucker observed score equating methods

	Total		Female		Male	
Raw Score	Equated Score	Difference	Equated Score	Difference	Equated Score	Difference
0	0.945	-0.945	0.722	-0.722	1.007	-1.007
1	1.936	-0.936	1.935	-0.935	2.021	-1.021
2	2.927	-0.927	2.822	-0.822	3.034	-1.034
3	3.917	-0.917	3.787	-0.787	4.048	-1.048
4	4.908	-0.908	4.752	-0.752	5.062	-1.062
5	5.898	-0.898	5.717	-0.717	6.075	-1.075
6	6.889	-0.889	6.682	-0.682	7.089	-1.089
7	7.879	-0.879	7.647	-0.647	8.102	-1.102
8	8.870	-0.870	8.612	-0.612	9.116	-1.116
9	9.860	-0.860	9.577	-0.577	10.129	-1.129
10	10.851	-0.851	10.542	-0.542	11.143	-1.143
11	11.841	-0.841	11.507	-0.507	12.157	-1.157
12	12.832	-0.832	12.472	-0.472	13.170	-1.170
13	13.822	-0.822	13.437	-0.437	14.184	-1.184
14	14.813	-0.813	14.401	-0.401	15.197	-1.197
15	15.803	-0.803	15.366	-0.366	16.211	-1.211
16	16.794	-0.794	16.331	-0.331	17.224	-1.224
17	17.784	-0.784	17.296	-0.296	18.238	-1.238
18	18.775	-0.775	18.261	-0.261	19.252	-1.252
19	19.765	-0.765	19.226	-0.226	20.265	-1.265
20	20.756	-0.756	20.191	-0.191	21.279	-1.279
21	21.747	-0.747	21.156	-0.156	22.292	-1.292
22	22.737	-0.737	22.121	-0.121	23.306	-1.306
23	23.728	-0.728	23.086	-0.086	24.319	-1.319
24	24.718	-0.718	24.051	-0.051	25.333	-1.333
25	25.709	-0.709	25.016	-0.016	26.347	-1.347
26	26.699	-0.699	25.981	0.019	27.360	-1.360
27	27.690	-0.690	26.946	0.054	28.374	-1.374
28	28.680	-0.680	27.911	0.089	29.387	-1.387
29	29.671	-0.671	28.876	0.124	30.401	-1.401
30	30.661	-0.661	30.661	-0.661	31.415	-1.415
31	31.652	-0.652	30.806	0.194	32.428	-1.428
32	32.642	-0.642	32.642	-0.642	33.442	-1.442
33	33.633	-0.633	33.632	-0.632	34.455	-1.455
34	34.622	-0.622	34.066	-0.066	35.469	-1.469
35	35.614	-0.614	35.613	-0.613	36.482	-1.482
36	36.603	-0.603	36.028	-0.028	37.496	-1.496

When Table 6 is examined, shown raw scores range between 0-36. It is seen that equated scores for all groups range between 0.945 and 36.603. For women range between 0.722-36.028 and for men range between 1.007-37.496. As can be seen from the table, according to Tucker equating method, the equated scores between 26-29 range and at 31st raw scores are smaller than the raw scores and the other equated scores are bigger than the raw scores. It was also found that in males, equated scores are higher than the raw scores. Based on these findings, it can be said that 6th booklet was more difficult than 4th one for whole-group and males. Although this was the case for females in a general sense, this situation changes between 26-29 interval and 31st raw scores.

Table 7. Raw Scores and the scores corresponding to the raw scores that are obtained via Levine observed score equating method

	Total		Female		Male	
Raw Score	Equated Score	Difference	Equated Score	Difference	Equated Score	Difference
0	1.167	-1.167	1.159	-1.159	1.000	-1.000
1	2.164	-1.164	2.164	-1.164	2.032	-1.032
2	3.155	-1.155	3.356	-1.356	3.063	-1.063
3	4.146	-1.146	4.293	-1.293	4.094	-1.094
4	5.137	-1.137	5.229	-1.229	5.125	-1.125
5	6.128	-1.128	6.166	-1.166	6.157	-1.157
6	7.118	-1.118	7.103	-1.103	7.188	-1.188
7	8.109	-1.109	8.039	-1.039	8.219	-1.219
8	9.100	-1.100	8.976	-0.976	9.251	-1.251
9	10.091	-1.091	9.912	-0.912	10.282	-1.282
10	11.082	-1.082	10.849	-0.849	11.313	-1.313
11	12.073	-1.073	11.785	-0.785	12.344	-1.344
12	13.064	-1.064	12.722	-0.722	13.376	-1.376
13	14.054	-1.054	13.659	-0.659	14.407	-1.407
14	15.045	-1.045	14.595	-0.595	15.438	-1.438
15	16.036	-1.036	15.532	-0.532	16.469	-1.469
16	17.027	-1.027	16.468	-0.468	17.501	-1.501
17	18.018	-1.018	17.405	-0.405	18.532	-1.532
18	18.775	-0.775	18.341	-0.341	19.563	-1.563
19	19.999	-0.999	19.278	-0.278	20.594	-1.594
20	20.990	-0.990	20.214	-0.214	21.626	-1.626
21	21.981	-0.981	21.151	-0.151	22.657	-1.657
22	22.972	-0.972	22.088	-0.088	23.688	-1.688
23	23.963	-0.963	23.024	-0.024	24.719	-1.719
24	24.954	-0.954	23.961	0.039	25.751	-1.751
25	25.944	-0.944	24.897	0.103	26.782	-1.782
26	26.935	-0.935	25.834	0.166	27.813	-1.813
27	27.926	-0.926	26.770	0.230	28.844	-1.844
28	28.917	-0.917	27.707	0.293	29.876	-1.876
29	29.908	-0.908	28.643	0.357	30.907	-1.907
30	30.899	-0.899	30.898	-0.898	31.938	-1.938
31	31.889	-0.889	30.517	0.483	32.969	-1.969
32	32.880	-0.880	32.880	-0.880	34.001	-2.001
33	33.871	-0.871	33.871	-0.871	35.032	-2.032
34	34.854	-0.854	33.948	0.052	36.064	-2.064
35	35.853	-0.853	35.852	-0.852	37.095	-2.095
36	36.836	-0.836	35.877	0.133	38.127	-2.127

As can be seen from Table 7, while the raw scores between 0-36 score interval, the equated scores change between 1.167 and 36.836 for the whole-group, 1.159-35.877 for females and 1-38.127 for

males. The results of the Levine observed score equating indicate that raw scores for whole-group and males are lower than the equated scores. However for females while raw scores are lower than equated scores between 0-23 raw score interval, they are lower and higher for some scores between 24-36 score interval.

In linear equating that regulates the difficulty difference of the forms across all scale scores, it was revealed that in both methods used in the study, there was a linear relationship between raw scores and equated scores for whole-group and males. There is no difference across whole number scale and only show difference between 24-36 score interval. It is fair to say that in Levine observed score equating 6th booklet was found to be more difficult than 4th one for whole-group and males and although it was the case for females in a general sense, this situation changes in raw scores between 24-31 interval.

The graphs of RMSD (x) index that correspond to each score in which group invariance of the Tucker and Levine observed score equating is examined according to the gender subgroup are presented in Figure 1 and Figure 2 respectively. The RMSD (x) values are given in the appendix in Table 1.

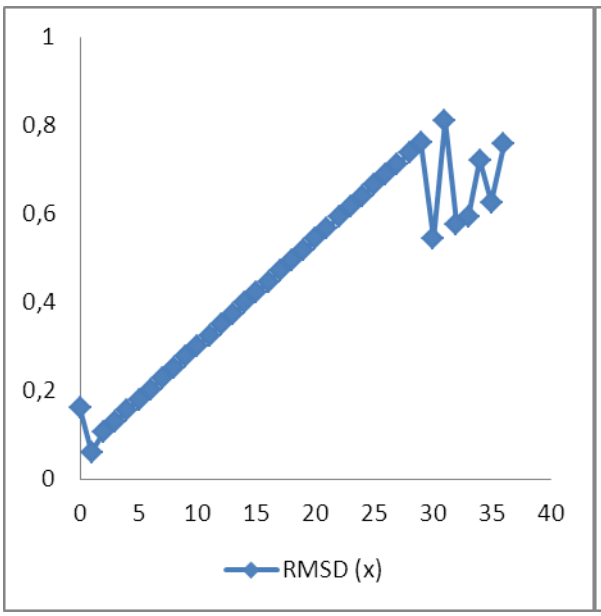


Figure 1. RMSD (x) for Tucker Equating

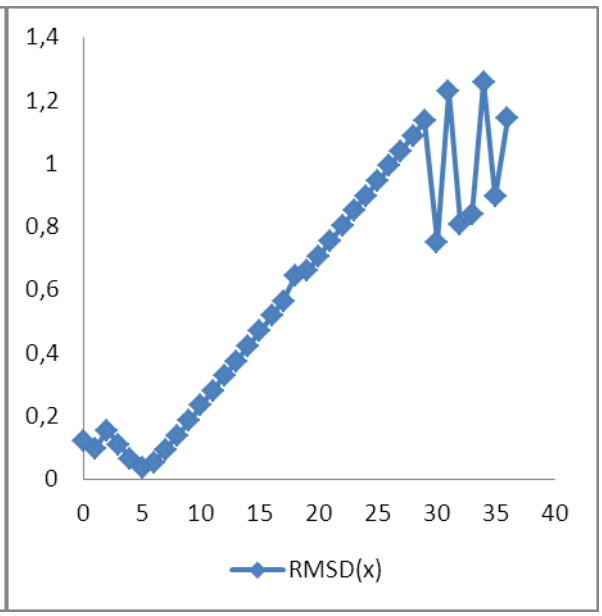


Figure 2. RMSD(x) for Levine Equating

When Figure 1 and Table 1 in appendix are examined, it is seen that RMSD (x) values range between 0.061 and 0.811 for Tucker equating and these values increased in simultaneously with the score in a general sense. However, this case differs when it comes to high scores. For Tucker equating method, the highest RMSD (x) value was obtained at 31 score level and the lowest one was obtained at 1 score level. In Figure 2, it is seen that RMSD (x) values for Levine Equating range between 0.034 and 1.257. Although it is seen that RMSD (x) values increased in simultaneously with the score in Levine equating method, it was found out that the increase was not linear at extreme values. In this method, the highest RMSD(x) score was obtained at 34 score level and the lowest one at 5 score level. According to RMSD values, there are some fluctuations in the extreme points of the scale in the graph for both equating methods. When the frequency of scores was examined, some extreme scores had fewer frequency than the others. Accordingly, fluctuations in the extreme points can be originated from the difference of frequencies.

In this study, it was found out that RMSD(x) values calculated with both methods were similar, however, RMSD(x) values for Tucker were smaller than the RMSD(x) values for Levine.

On the other hand, it is seen that RMSD(x) values that correspond to the scores between 1 and 18 for Tucker equating are lower than DTM. This means that the difference between the equated score in

whole-group and equated scores in sub-groups is not significant. However, RMSD(x) values that correspond to the scores between 19 and 35 for Tucker equating are higher than DTM which means that the difference between equated score in whole group and equated scores in sub-groups is significant. For Levine equating, it is seen that RMSD(x) values that correspond to the scores between 1 and 15 for are lower than DTM. This means that the difference between the equated score in whole-group and equated scores in sub-groups is not significant. However, RMSD(x) values that correspond to the scores between 16 and 35 are higher than DTM. Therefore the difference between equated scores in whole-group and equated scores in sub-groups is significant.

RMSD (x) index that correspond to each score in which group invariance of the scores equated according to Tucker and Levine observed score equating in gender sub-group is examined is given above. REMSD values that are calculated at group invariance total score level are presented in Table 8.

Table 8. Values for Levine and Tucker Equating Methods

Equating Methods	REMSD
Tucker-Linear Equating	0.496
Levine-Linear Equating	0.668

As shown in the Table 8 Tucker equating, REMSD value was calculated as 0.496 and as 0.668 for Levine equating method. It is seen that REMSD value obtained for Tucker is lower than the REMSD value obtained for Levine. Besides Tucker equating RMSD(x) values are lower than DTM. This implies the difference between the equated score in whole-group and equated scores in sub-groups is not significant. However, for Levine equating, it is seen that RMSD(x) values are higher than DTM. This means that the difference between equated scores in whole-group and equated scores in sub-groups is significant.

WMSE (AHKO) coefficients were calculated according to Tucker and Levine equating methods and gender sub-group determined for invariance in order to find if Tucker or Levine is more suitable for the PISA 2012 4th and 6th booklets which included mathematics test. The information regarding coefficients is presented in Table 9.

Table 9. WMSE (AKHO) Values for Levine and Tucker Equating Methods

Equating Methods	Total	Female	Male
Tucker-Linear Equating	0.012	0.004	0.024
Levine-Linear Equating	0.199	0.112	0.035

Table 9 indicates that according to whole-group and sub-groups, the most suitable method regarding the mathematics sub-test in PISA 2012 included in 4th and 6th booklets is Tucker equating method. It is striking that in Tucker equating method, WMSE value obtained for males is quite higher than the WMSE value obtained for females. It is fair to say that WMSE coefficients obtained for males via both methods from sub-groups are similar.

DISCUSSION and CONCLUSION

In this study, equating errors of the scores obtained according to Tucker and Levine observed score equating methods were compared by equating with the 6th and 4th booklets of PISA 2012 Mathematics

subtest and in order to assess the equitability of the scores, whether or not group invariance is was investigated according to RMSD (x) for each score and REMSD coefficients for total score.

When the scores obtained via linear equating are examined, it was seen that the scores obtained according to Tucker and Levine observed score equating take values out of raw score range. Livingston (2004) maintained that the scores equated in linear equating can go outside the raw score range and that does not create a problem for linear equating and is a characteristics specific to linear equating. Moreover, Livingston (2004) reported that the equated scores at very high and low scores can exceed the score range. This was observed at high scores in both equating methods according to female sub-group.

When WMSE values obtained based on the Tucker and Levine observed score equating methods are compared, it was found out that Tucker observed score equating produced lowest error for both whole-group and gender sub-group. While errors that are obtained according to Tucker and Levine observed score equating with regard to whole-group and female sub-group show difference, it can be said that errors that are obtained with regard to males sub-group are close. Similar results are obtained when the past studies are examined. A study by Demir & Güler (2014) compared frequency prediction equipercentile equating, Tucker, Levine and Braun-Holland Linear Equating methods and determined that the most appropriate method was Tucker equating method and also reported that Levine observed score equated method had the highest error. Topczewski et al., (2013) stated in their study in which they used a different version of Tucker, Angoff-Levine, congeneric -Levine and a different version of congeneric Levine by addressing the differences between the skills of the groups that Tucker equating method was the most suitable one in case that group variance is similar. Chen et al. (2003) performed Tucker and Levine observed score equating methods by using different skills distribution and tests with different difficulty levels and concluded that the results were similar when the difference between the group and tests forms was small.

When RMSD and REMSD values obtained according to gender sub-group via linear equating are examined, it was seen that the RMSD and REMSD values based on Tucker were lower than the ones based on Levine. Besides, the difference between the equated scores in whole-group and the scores equated for sub-groups is not significant for Tucker equating method, although it is significant for Levine equating method. That is to say that while group invariance is at an acceptable level for Tucker equating method, it is not the case for Levine equating method. In the study by von Davier and Han (2004) which compared RMSD values with respect to gender with Levine observed score and chained linear equating methods, it was observed that the equating function with the lowest changing equating rate belonged to Levine while the highest changing function belonged to Tucker method. It was found out that the present study and the relevant study results were not parallel. The study by Dorans, Liu & Hammond (2008) reported in their study in which they compared group invariance by gender with Tucker, Levine and Chained equating methods revealed that if the groups to be equated are similar in terms of average skills, Trucker equating method is more fruitful than Levine and Chained equation results. Also Yin, Brennan & Kolen (2004) investigated the group invariance of linear, parallel-linear and equipercentile equating of mathematics and science tests in their study. They reported that lower REMSD values were obtained via linear and parallel linear equating methods for mathematics tests, while lower REMSD via equipercentile equating was reported for the science test. It is seen that results of both studies support the current study.

Equitability of scores requires the same meaning regardless of when or when the equalized points are applied. Failure to achieve group invariance in equating function indicates that the difficulty difference of the old and new test forms in NEAT pattern is inconsistent across subgroups (Kim and Walker, 2009). Violation of group invariance condition in equating causes the individuals from different groups who are supposed to have the same score get different equated scores (Dorans, 2004, 2008). Group invariance is a prerequisite for equating. Failure to achieve group invariance is an indicator that equating has not succeeded completely. However, achieving group invariance does not necessarily mean that equated scores can be used interchangeably. This is because group invariance should not be taken as the only criterion in assessing the quality of the equation (Dorans, Liu & Hammond, 2008).

The usage of group invariance indexes made it possible to decide which equating method can be achieved better than the other. Based on the findings of current study, Tucker equating method was the best option in terms of equating 4th and 6th Mathematics Booklets of PISA 2012 and group invariance.

The difference observed in group invariance might be attributed to the difference between the whole and sub-group samples. The sample size of this study is 741, 381 and 360 for the whole group, males and females, respectively and a sample size between 50 and 100 is sufficient for Tucker and Levine observed score equating methods (Parshall, Du Bose Houghton & Kromrey, 1995; Skaggs, 2005; Babcock, Albano & Raymond, 2012). Since the sample sizes are sufficient in this study, it can be said that the difference in group invariance is not affected by the sample size.

In this study, 4th and 6th booklets of PISA 2012 mathematics sub test were equated by using Tucker and Levine observed score equating method in non-equivalent groups' common items pattern and it was investigated whether or not group invariance was achieved with regard to gender sub-group. A similar study can be carried out by using different equating methods, equating patterns and different samples. Also, whether or not group invariance condition was met with regard to gender sub-group was examined via RMSD (x) and REMSD indexes. In different studies, difference group invariance indexes can be used according to different sub-groups (socioeconomics, ethnic groups, countries etc.).

REFERENCES

Akhun, İ. (1984). İki korelasyon katsayısı arasındaki manidarlığın test edilmesi. *Ankara Üniversitesi Eğitim Fakültesi Dergisi. 17*, 1-7.

Angoff, W. (1996). Scales, norms, and equivalent scores. *Educational Measurement: Theories and Applications*, *2*, 121.

Atalay Kabasakal, K. (2014). *Değişen madde fonksiyonunun test eşitlemeye etkisi* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Baykul, Y. (1996*). İstatistik: Metodlar ve uygulamalar* (3. Baskı). Ankara: Anı Yayıncılık

Babcock, B., Albano, A., & Raymond, M. (2012). Nominal weights mean equating: A method for very small samples. *Educational And Psychological Measurement, 72*(4), 608-628.

Bozdağ, S., & Kan, A. (2010). Şans başarısının test eşitlemeye etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 39*, 91-108.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. *Test equating*, 1982, 9-49.

Chen, H., Cui, Z., Zhu, R., & Gao, X. (2010). *Evaluating the effects of differences in group abilities on the Tucker and the Levine observed-score methods for common-item nonequivalent groups equating.* ACT Research Report Series (2010-(1)). Iowa City, IA: ACT.

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

Demir, S., & Güler, N. (2014). Ortak maddeli denk olmayan gruplar desenine ilişkin test eşitleme çalışması. *International Journal of Human Sciences, 11*(2), 190-208.

Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*, 43-68.

Dorans, N. J. (2008). *Three facets of fairness*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Measurement, 37*, 281-306.

Dorans, N. J., Holland, P. W., Thayer, D. T., & Tateneni, K. (2003). Invariance of score linking across gender group for three Advanced Placement Program examinations. In N. J. Dorans (Ed.). *Population invariance o f score linking: Theory and applications to Advanced Placement Program examinations* (RR-03-27). Princeton, NJ: Educational Testing Service.

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*, 81-98.

Gök, B. (2012). *Denk olmayan gruplarda ortak madde deseni kullanılarak Madde Tepki Kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer.

İnci, Y. (2014). *Örneklem büyüklüğünün test eşitlemeye etkisi* (Yayınlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Kahraman, H. (2012). *Düzgünleştirilmiş puanların eşitleme hatasına etkisi* (Yayınlanmamış Yüksek Lisans Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Kan, A. (2011). Test eşitleme: OKS testlerinin istatistiksel eşitliğinin sınanması. *Eğitim ve Bilim, 36*(160), 38-51.

Kelecioğlu, H. (1994). *Öğrenci Seçme Sınavı puanlarının eşitlenmesi üzerine bir çalışma* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü Ankara.

Kelecioğlu, H., & Öztürk Gübeş, N. (2013). Comparing linear equating and equipercentile equating methods using random groups design. *International. Online Journal of Educational Sciences, 5*(1), 227-241.

Kilmen, S. (2010). *Madde Tepki Kuramına dayalı test eşitleme yöntemlerinden kestirilen eşitleme hatalarının örneklem büyüklüğü ve yetenek dağılımına göre karşılaştırılması* (Yayınlanmamış doktora tezi). Ankara Üniversitesi, Eğitim Bilimler Enstitüsü, Ankara.

Kim, S., & Walker, M. E. (2009). *Evaluating subpopulation invariance of linking functions to determine the anchor composition for a mixed-format test*. ETS Research Rep. No. RR-09-36. Princeton, NJ: Educational Testing Service.

Kolen, M. J. (1988). An NCME intructional module on traditional equating methodology. *Educational Measurement: Isuues and Practice, 7*, 29-36.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*, 3-14.

Kolen, M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd Ed.). New York: Springer.

Kolen, M., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd Ed.). New York: Springer.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability*. ETS Research Report Series, 1955(2).

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Service.

MEB. (2010). *PISA 2009 projesi ulusal ön raporu*. MEB Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı.

MEB. (2013). *PISA 2012 ulusal ön raporu*. MEB Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

Mutluer, C. (2013). *Yıl içinde farklı dönemlerde yapılan Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) puanlarına ilişkin bir test eşitleme çalışması* (Yayınlanmamış yüksek lisans tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

OECD. (2014). PISA Technical Report, OECD Publishing. http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Öztürk, N. (2010). *Akademik personel ve lisansüstü eğitimi giriş sınavı puanlarının eşitlenmesi üzerine bir çalışma* (Yayınlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü. Ankara.

Öztürk Gübeş, N., & Kelecioğlu, H. (2017). Investigating group invariance of equating results. *Elementary Education Online, 16*(1), 217-227.

Parshall, C. G., Du Bose Houghton, P., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32*, 37-54.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.) *Educational Measurement* (pp.221-262). New York: Macmillan.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495–529.

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12*, 69–82.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*, 309-330.

Şahhüseyinoğlu, D. (2005). *İngilizce yeterlik sınavı puanlarının üç farklı eşitleme yöntemine göre karşılaştırılması* (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü. Ankara.

Topczewski, A., Cui, Z., Woodruff, D., Chen, H., & Fang, Y. (2013). *Comparison of four linear equating methods for the common-item nonequivalent groups design using simulation methods*. ACT Research Report Series (2013-(2). Iowa City, IA: ACT.

Uysal, İ. (2012). *Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması* (Yayınlanmamış yüksek lisans tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

von Davier, A. A., & Han, N. (2004). *Population invariance and linear equating for the non-equivalent groups design*. (ETS RR-04-47). Princeton, NJ: Educational Testing Service.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating.* New York: Springer.

von Davier, A. A., & Kong, N. (2003). *A unified approach to linear equating for the non-equivalent groups design*. (ETS SR-03-31). Princeton, NJ: Educational Testing Service.

von Davier, A. A., & Wilson, C. (2008). Investigating the population sensitivity assumption of Item Response Theory true-score equating across two subgroups of examinees and two test formats. *Applied Psychological Measurement, 32*, 11-26.

Yang, W. L., & Gao, R. (2008). Invariance of score linkings across gender groups for forms of a testlet-based college-level examination program examination. *Applied Psychological Measurement, 32*, 45-61.

Yang, W. L. (2004). Sensitivity of linkings between AP multiple-choice scores and composite scores to geographical region: An illustration of checking for population invariance. *Journal of Educational Measurement, 41*, 33-41.

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and ITED scores from different populations. *Applied Psychological Measurement, 28*, 274-289.

Yi, Q., Harris, D. J., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement, 32*, 62-80.

UZUN ÖZET

Giriş

Eşitleme benzer içerik ve güçlük düzeyinde geliştirilen test formları arasındaki farklılıkları düzenleyerek, bu formlardan elde edilen puanların birbiri yerine kullanılmasını sağlayan istatistiksel bir süreç olarak tanımlanabilir. Test eşitlemede amaç, kolay ya da zor test formunu alan bireye formun herhangi bir avantaj veya dezavantaj sağlamamasıdır. Test formlarının eşitlenebilmesi için eşitlik, simetri, grup değişmezliği ve tek boyutluluk gibi bazı koşulların karşılanması gerekmektedir. Bu koşullardan biri olan grup değişmezliği, eşitleme fonksiyonunun alt gruplardan bağımsız olması ve alt grupların eşitlemeyi etkilememesi anlamına gelmektedir. Bu araştırmanın amacı denk olmayan gruplarda ortak madde desenine göre doğrusal eşitleme yöntemlerinden Tucker ve Levine gözlenen puan eşitleme yöntemleriyle eşitlenmesi sonucunda cinsiyet alt grubuna göre eşitleme yöntemlerinin grup değişmezliği koşulunun sağlanıp sağlanmadığının incelenmesidir.

Yöntem

Bu araştırmanın örneklemini Türkiye'deki PISA 2012 uygulamasına katılan öğrenciler arasından, bu uygulama esnasında 4. ve 6. kitapçıkları alan 741 öğrenci oluşturmaktadır. Bu çalışmada veri toplama aracı için PISA 2012 uygulanmasındaki 4 ve 6 nolu kitapçıklarda yer alan maddeler kullanılmıştır.

Bu araştırmada verilerin analizi dört aşamada gerçekleştirilmiştir. Verilerin analizinin birinci aşamasında, eşitleme koşullarının sağlanıp sağlanmadığı test edilmiştir.

Bunun için ilk olarak, verinin tek boyutlu olup olmadığı test edilmiştir. Tek boyutluluğun belirlenmesi için iki kategorili veriler için kullanılan tetrakorik korelasyona dayalı temel bileşenler faktör analizi yöntemi seçilmiştir. Formların ortalama güçlükleri arasında anlamlı bir farkın olup olmadığını belirlemek için iki oran fark testi yapılmıştır. Eşitlenecek kitapçıkların eşit güvenirliğe sahip olup olmadığını görebilmek için KR-20 güvenirlik katsayısı hesaplanmıştır. İki güvenirlik katsayısı arasında fark olup olmadığı belirlemek için Fischer'ın Z istatistiği yapılmıştır. Eşitleme yapılacak kitapçıkların ortalama ve varyansları arasında fark olup olmadığı bağımsız gruplar t testi ve Levene testi ile incelenmiştir. Eşitleme için gerekli koşullar ile ilgili yapılan analizler sonucunda, testlerin tek boyutlu olduğu; güvenirliklerinin, varyanslarının ve ortalama güçlüklerinin eşit olduğu görülmüştür.

Veri analizinin ikinci aşamasında, eşitleme yöntemleri kullanılarak eşitlenmiş puanlar elde edilmiştir.

Veri analizinin üçüncü aşamasında, her bir eşitleme yöntemi ile elde edilen eşitlenmiş puanların cinsiyete göre nasıl değiştiğini değerlendirmek için grup değişmezliği indeksleri hesaplanmıştır. Bu çalışmada grup değişmezliğini belirlemek için Dorans ve Holland (2000) tarafından geliştirilen RMSD ve REMSD indeksleri kullanılmıştır. Eşitlemede grup değişmezliğinin değerlendirilmesinde DTM ölçütünden yararlanılmıştır. Bu çalışmada DTM= 0.50 kriteri alınarak bir puanın toplam gruptaki bir eşitlenmiş puan ile alt grup(lar)daki eşitlenmiş puan(lar) arasındaki farklılığın 0.50'den daha az

olmasının yok sayılabilir; 0.50'den daha fazla olmasının ise anlamlı olduğu kabul edilerek yorumlar yapılmıştır.

Veri analizinin son olarak dördüncü aşamasında, her bir eşitleme yönteminde yapılan hata miktarı hesaplanmıştır. Eşitleme hatasını değerlendirmek için ağırlıklandırılmış hata kareleri ortalaması (WMSE) ölçütü kullanılmıştır.

Sonuçlar ve Tartışma

Doğrusal eşitleme sonucunda elde edilen puanlar incelendiğinde, Tucker ve Levine gözlenen puan eşitlenmesine göre elde edilen puanların, ham puan ranjının dışında değerler aldığı görülmüştür.

Tucker ve Levine gözlenen puan eşitleme yöntemlerine dayalı olarak elde edilen WMSE değerleri karşılaştırıldığında ise, hem tüm grup hem de cinsiyet alt grubuna göre en az hata veren yöntemin Tucker gözlenen puan eşitleme olduğu görülmüştür. Toplam grup ve kadın alt grubuna göre Tucker ve Levine gözlenen puan eşitlemesine göre elde edilen hata değerleri büyük farklılık gösterirken, erkekler alt grubuna göre elde edilen hata değerlerinin birbirine yakın olduğu bulunmuştur.

Doğrusal eşitleme ile cinsiyet alt grubuna göre elde edilen RMSD ve REMSD değerleri incelendiğinde, Tucker için elde edilen RMSD ve REMSD değerlerinin Levine için elde edilen değerlerden daha küçük olduğu görülmüştür. Ayrıca toplam gruptaki eşitlemiş puan ile alt gruplardaki eşitlenmiş puanlar arasındaki farklılığın Tucker eşitleme yöntemi için anlamsız olduğu; Levine eşitleme yöntemi için ise anlamlı olduğu sonucuna ulaşılmıştır.

Puanların eşitlenebilirliği, eşitlenmiş puanların ne zaman ya da hangi gruba uygulandığına bakılmaksızın aynı anlama gelmesini gerektirmektedir. Eşitleme fonksiyonundaki grup değişmezliğinin sağlanamaması, NEAT deseninde eski ve yeni test formlarının güçlüklerindeki farklılığın alt gruplar boyunca tutarlı olmadığını göstermektedir. Eşitlemede grup değişmezliğinin ihlali, aynı puana sahip olması gereken farklı gruplara ait bireylerin, farklı eşitlenmiş puanlar almasına neden olmaktadır. Grup değişmezliği eşitleme için bir önkoşuldur. Grup değişmezliğinin sağlanamaması eşitlemenin tam olarak gerçekleşmediğinin kanıtı olarak ele alınabilir. Ancak grup değişmezliğinin sağlanması da eşitlenmiş puanların birbiri yerine kullanılabileceği anlamına gelmez. Çünkü, eşitlemenin niteliğinin değerlendirilmesinde tek kriter grup değişmezliği değildir.

Bu çalışmada grup değişmezliği indekslerinin kullanımı, puanların eşitlenebilirliğinin hangi yöntemle daha iyi sağlandığına karar verilmesine olanak vermiştir. Elde edilen bulgulara dayalı olarak, PISA 2012 matematik 4. ve 6. Kitapçıkların eşitlenmesinde ve grup değişmezliği açısından en uygun yöntemin Tucker eşitleme yöntemi olduğuna ulaşılmıştır.

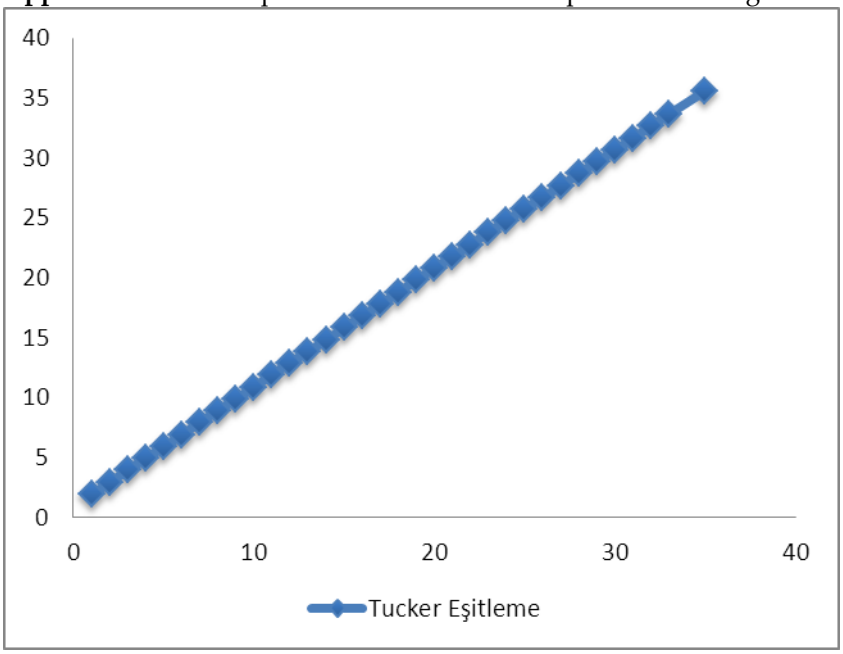
Bu araştırmada PISA 2012 matematik 4. ve 6. Kitapçıklar denk olmayan gruplarda ortak madde test deseninde Tucker ve Levine gözlenen puan eşitleme yöntemi kullanılarak eşitlenmiş ve cinsiyet alt grubuna göre grup değişmezliğinin sağlanıp sağlanamadığı incelenmiştir. Benzer bir araştırma farklı eşitleme yöntemleri, eşitleme desenleri ve farklı örneklemler kullanılarak yapılabilir. Ayrıca grup değişmezliğinin sağlanıp sağlanamadığı cinsiyet alt grubuna göre RMSD ve REMSD indeksleriyle yapılmıştır. Diğer çalışmalarda farklı alt gruplara (sosyo ekonomik, etnik grup, ülkeler vb.) göre farklı grup değişmezliği indeksleri ile yapılabilir.

APPENDICES

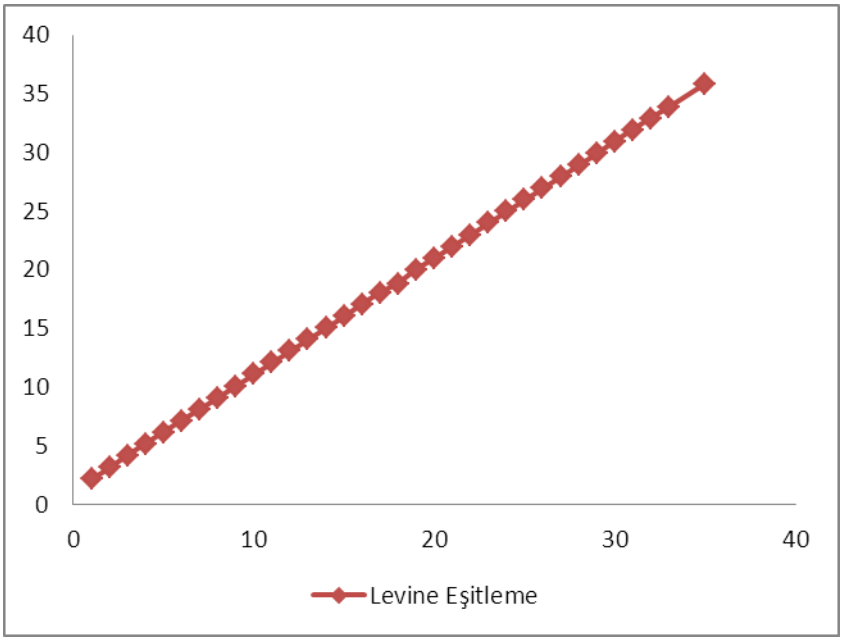
Appendix 1. RMSD (x) Values Regard to Equating Methods

Raw Score	Tucker Linear Equating	Levine Linear Equating
0	0,161	0,121
1	0,061	0,096
2	0,106	0,154
3	0,131	0,108
4	0,155	0,065
5	0,179	0,034
6	0,203	0,051
7	0,228	0,093
8	0,252	0,139
9	0,276	0,185
10	0,300	0,232
11	0,325	0,279
12	0,349	0,327
13	0,373	0,374
14	0,398	0,421
15	0,422	0,469
16	0,446	0,516
17	0,470	0,564
18	0,495	0,643
19	0,519	0,659
20	0,543	0,706
21	0,568	0,754
22	0,592	0,801
23	0,616	0,849
24	0,641	0,896
25	0,665	0,944
26	0,689	0,992
27	0,713	1,039
28	0,738	1,087
29	0,762	1,134
30	0,543	0,750
31	0,811	1,229
32	0,576	0,808
33	0,593	0,837
34	0,722	1,257
35	0,626	0,895
36	0,757	1,144

Appendix 2. The Graphics of Raw Score and Equated score Regard to Equating Methods



The Graphic of Raw Score and Equated score Regard to Tucker Linear Equating Method



The Graphic of Raw Score and Equated score Regard to Levine Linear Equating Method

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);146-157

Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu: Geçerlik ve Güvenirlik Çalışmaları

University Students Form of Emotional Jealousy Scale: Validity and Reliability Studies

Seval KIZILDAĞ*

Öz

Bu çalışmanın amacı, Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu'nun geçerlik ve güvenirlik çalışmalarını yapmaktır. Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu üç faktör (değersizlik hissi; ilişkisel doyumsuzluk ve aşkın yitimi ve beraber zaman geçirmede isteksizlik) ve 17 maddeden oluşmaktadır. Araştırmanın iki çalışma grubu bulunmaktadır. Birinci çalışma grubunu 158'si kadın (%37) ve 92'si erkek (%37) toplam 250 üniversite öğrencisi; ikinci çalışma grubunu da test tekrar test güvenirlik analizleri için çalışmaya dahil edilen 18'si kadın (%54.5) ve 15'i erkek (%45.5) toplam 33 üniversite öğrencisi oluşturmaktadır. Güvenirlik analizleri kapsamında Cronbach alfa güvenirlik katsayısı, madde toplam korelasyonları ve test-tekrar test güvenirliği incelenmiştir. Ölçeğin Cronbach alfa değeri .94; test-tekrar test korelasyon katsayısı .97 bulunmuştur. Ölçek madde toplam korelasyon değerlerinin .43 ile .79 arasında değiştiği gözlenmektedir. Geçerlik analizi kapsamında, yapı geçerliği ve yordama geçerliğine bakılmıştır. Doğrulayıcı faktör analizleri sonucunda üç faktörlü model incelendiğinde χ^2=334.29, χ^2/df=2.88; RMSEA=.09, NFI=.88, CFI=.92, IFI=.92, TLI=.91 ve RFI=.86 değerleri elde edilmiştir. Bu değerler modelin veri uygunluğunun yeterli olduğunu göstermektedir. Özetle, ölçeğin üniversite öğrencileri üzerinde yapılan geçerlik ve güvenirlik çalışmaları sonuçları, bu ölçeğin üniversite öğrencilerinin kıskançlık düzeyini ölçebilecek niteliklere sahip olduğunu göstermektedir.

Anahtar Kelimeler: romantik kıskançlık, duygusal kıskançlık ölçeği, üniversite öğrencileri

Abstract

The aim of this study is to conduct the validity and the reliability studies for University Student Form of Emotional Jealousy Scale. University Student Form of Emotional Jealousy Scale is a scale that consists of 17 items and three factors as "feeling of valueless, relational dissatisfaction and loss of love, unwillingness for having time together". There are two study groups in this research. First study group consists of 158 female students (63%) and 92 male students (37%) and second study group consists of 18 female students (54.5%) and 15 male students (45.5%) for test-retest reliability analysis. Cronbach alpha reliability coefficient, total-item correlation values, and test-retest reliability were examined on the scope of reliability analysis. Cronbach alpha value of the scale was found as .94; and test-retest correlation coefficient was found as .97. It is observed that total-item correlation values of the scale change between .43 and .79. Confirmatory factor analysis, predictive validity, and findings directed to reliability studies concerning construct validity of the scale were presented. Model fit indices were found as χ^2= 334.29, χ^2/df=2.88; RMSEA=.09, NFI=.88, CFI=.92, IFI=.92, TLI=.91, and RFI=.86 in the wake of analysis. These values show that model-data fit is sufficient. To sum up, it is concluded with the analysis that the university form of the scale is a valid and reliable measurement tool.

Keywords: romantic jealousy, emotional jealousy scale, university students

GİRİŞ

Romantik kıskançlık, bireyin kendilik değerine veya ilişkisine yönelik tehditler içeren karmaşık duygu, düşünce ve davranışlar bütünü olarak tanımlanmaktadır (White, 1981). Pines (1992a) ise romantik kıskançlığı, değer verilen bir ilişkiye veya onun niteliğine yönelik tehdit algısıyla oluşan karmaşık bir tepki olarak tanımlamakta ve romantik kıskançlığın içsel ve dışsal olmak üzere iki

* Arş. Gör. Dr., Adıyaman Üniversitesi, Eğitim Fakültesi, Adıyaman-Türkiye, e-posta: sevalpdr@gmail.com

DOI: *10.21031/epod.302673*
Geliş Tarihi: 25.11.2016
Kabul Tarihi: 24.03.2017

boyuttan oluştuğunu ifade etmektedir. Romantik kıskançlığın içsel boyutunda, belli duygular, bilişler ve fiziksel semptomların; dışsal boyutunda ise daha kolay bir şekilde dış dünyadan fark edilebilir davranışların yer aldığını belirtmektedir. Romantik kıskançlık tanımlarında ortak olan durum, sevilen birinin kaybedilme korkusuna ilişkin diğer kişinin yaşadığı karmaşık duygu, düşünce ve davranışları içermesidir. Bir başka ifadeyle, romantik kıskançlık çok boyutlu ve kapsamlı bir psikolojik yapıdır.

Literatürde cinsel kıskançlık ve duygusal kıskançlık olarak da sınıflandırılabilen romantik kıskançlığın her iki türünde de ortak olan davranışlar; çiftlerden birinin ilişki için daha az çaba harcaması ve partneri (eşi/sevgilisi) için çekici ve arzu edilebilir olmasındaki isteksizlik halinde birleşmektedir (Shackelford ve Buss, 1997). Buna karşın, çiftlerden birinde ilişki dışındaki üçüncü kişiyle daha fazla iletişim kurma, zaman geçirmeye isteklilik görülebilir. Üçüncü kişiye yönelik bu eğilim ile belirli bir başlatıcı olay arasındaki etkileşim romantik kıskançlığın bir sonucudur fakat kıskançlığı ortaya çıkarıcı bir olay yaşanmadığı sürece bu duygunun yaşanma olasılığı da düşebilir (Pines,1992b). Başlatıcı olay bazen partnerin bir başka kişiye ilgisinden dolayı kendi eşinden uzaklaşması ve ayrılması korkusuyla da tetiklenmektedir (Sharpsteen ve Kirkpatrick, 1997). Buradan hareketle, cinsel veya duygusal kıskançlık durumlarında temelde sevilen kişiyi bir şekilde kaybetme korkusunun öne çıktığı söylenebilir.

Romantik kıskançlığın boyutlarından biri olan cinsel kıskançlık, bireyin eşinin bir başkasıyla cinsel beraberlik yaşadığını bilmesi ya da bundan şüphelenmesi sonucunda yaşanan kıskançlıktır (Demirtaş-Madran, 2008). Shackelford ve Buss (1997) cinsel kıskançlığın en belirgin özelliklerini bazen cinsel isteksizlik veya uzaklaşma bazen abartılı bir şekilde cinsel konularla ilgilenme veya konuşma gibi tutarsız davranışlarla ilişkilendirmektedir. Aynı zamanda, cinsel davranışlarda normalden farklı davranma (sıklık veya yaşanma şekli gibi) cinsel kıskançlığın belirleyicileri olabilmektedir. Görüleceği üzere, romantik ilişkideki bir davranışın normal seyrinin dışında yaşanması ve çiftlerden birinin bu durumu tehdit olarak algılaması ve nihayetinde de romantik kıskançlığı yaşamasına neden olabilir.

Romantik kıskançlığın bir diğer boyutu olan ve bu çalışmanın temel alt yapısını oluşturan duygusal kıskançlığı ortaya çıkaran davranışlar ise Shackelford ve Buss (1997) tarafından ilişkisel doyumsuzluk ve aşkın yitimi; duygusal ihmal; beraber zaman geçirmede isteksizlik; pasif reddetme ve düşüncesizce davranışlar sergilemeye başlama; öfkeli, eleştiriye dayalı ve sorgulayıcı iletişime girme; belirli bir birey hakkında konuşmaktan kaçınma ve suçlu ve kaygılı bir iletişim tarzı benimseme olarak sıralanmıştır. Bu davranışların temelindeki duyguların düzenlenmesi veya kontrol edilmesi, romantik ilişkinin olumlu sürekliliği açısından önemli görülmektedir. Duyguların düzenlenmesinde yaşanan zorluklar duygusal tepkiyi kabul etmeme, amaç yönelimli davranışları sürdürmekte ve dürtü kontrolünde zorlanma, duygu düzenleme stratejilerini sınırlı kullanma, duygusal farkındalıktan yoksunluk ile ilişkilendirilmektedir (Gratz ve Roemer, 2004). Buradan hareketle, duygusal kıskançlığın da olumsuz duygular yaşama veya onları kontrol edememe gibi olası sorunlar doğuracağı söylenebilir.

İlgili literatürde, romantik kıskançlık ilişkilerde genellikle olumsuz duygularla eşleştirilmektedir. Pfeiffer ve Wong (1989) duygusal kıskançlığı korku, öfke, güvensizlik, üzüntü gibi hislerle ilişkilendirirken; Wreen (1989) kıskançlığı güvensizlik, kaybetme korkusu ve duygusal soğukluk ile ilişkilendirmektedir. Romantik ilişkide korku, öfke, güvensizlik veya duygusal soğukluk gibi duyguların ilişkinin niteliğini de olumsuz etkileyebileceği söylenebilir. Sağlıklı bir ilişki güvensizlik, korku, ilgisizlik, değersizlik gibi durumların aksine; güven, ilgili olma, değerli hissetme gibi duygu durumlarını içermektedir. Bu sebeple, ilişkide sağlıklı bağlanma gerçekleştirmiş bireylerin romantik kıskançlıklarının da daha düşük düzeylerde olması beklenir.

Sharpsteen ve Kirkpatrick (1997) bağlanma ile romantik kıskançlık arasındaki ortak özellikleri ilişkiyi devam ettiren bir süreç olması; sevilen birinden gerçek veya olası ayrılmanın başlatılması; öfke, korku, üzüntü gibi duyguları içermesi ve kendiliğin ve ilişkilerin zihinsel modeli tarafından düzenlenmesi açılarından ortaya koymuşlardır. Buunk (1997) yüksek düzeyde kıskançlığın kaygılı bağlanan bireylerde, daha sonra kaçınan bireylerde gözlendiğini belirtirken; güvenli bağlanan bireylerde kıskançlığın daha düşük düzeyde olduğunu belirtmektedir. Bununla birlikte, romantik kıskançlığın kaygılı bağlanmayla ilişkili olduğu (Knobloch, Solomon ve Cruz, 2001) ve güvenli

bağlanma stiline sahip kişilerin duygusal aldatmayı cinsel aldatmadan daha stresli buldukları, kayıtsız bağlanma stiline sahip kişilerde ise cinsel aldatmayı daha büyük bir sorun olarak algıladıkları ifade edilmektedir (Levy ve Kelly, 2006). Görülebileceği üzere, ilişkilerde güven duygusu romantik kıskançlığın da yaşanma olasılığını ve biçimini etkileyebilmektedir.

Selterman ve Maier (2013) de güvenli bağlanan bireylerin güvensiz bağlanan bireylerden daha çok kıskançlık duygusunu deneyimlediklerini ve güvensiz bağlanmanın ilişkide tehdit yaratan yanlış algılamalarla ilişkilendirildiğini belirtmektedir. Simpson (1990) da güvensiz bağlanmada partnerin davranışına yönelik olumsuz davranış ve ihanet beklentilerinin arttığını ifade etmektedir. Aynı zamanda güvensiz bağlanan kişiler olumlu duygularını açıkça ifade edemezlerken; kızgınlık, öfke ve benzeri olumsuz duygularını ise düşmanca bir şekilde ifade edebilirler. Dolayısıyla güvensiz bağlanan kişilerde kıskançlık duygusunun ilişkiye daha zarar veren biçimde saldırganlık, şiddet gibi davranışa dönüşme olasılığı daha yüksektir. Bununla birlikte, güvensiz bağlanan bireylerin olumlu deneyimlerden daha az bahsettikleri ve bu olumlu deneyimlerin etkilerinden yeterince yararlanmadıkları da bilinmektedir (Gentzler, Kerns ve Keener, 2010). Bir başka ifadeyle, güvensiz bağlanan bireyler yaşamlarını daha çok olumsuz yaşantılar üzerinden anlamlandırmaktadırlar.

Edalati ve Redzuan (2010) sevginin azalması, reddetme, güvensizlik, kendine güvenin kaybedilmesi, duygusal destek ve değerli olma duygusunun azalması ve korku temelli romantik kıskançlık ile saldırganlık arasında güçlü bir ilişkinin olduğunu belirtmektedir. Örneğin, romantik ilişkisi olan üniversite öğrencilerinde bağlanma ve romantik kıskançlık ilişkisinin incelendiği Karakurt (2001) tarafından yapılan bir araştırmada, korkulu bağlanan bireylerin güvenli bağlananlardan daha yüksek düzeyde davranışsal kıskançlık gösterdikleri; saplantılı bağlananların da güvenli bağlananlardan daha yüksek düzeyde olumsuz ve yetersizlik duygusu taşıdıkları bulunmuştur. Aynı zamanda, söz konusu bu çalışmada romantik kıskançlıkla baş etmede, güvenli bağlanan bireylerin ilişkiyi korumaya daha yatkın oldukları; kayıtsız bağlananların ise daha düşük düzeyde ilişkiyi korumaya yönelik oldukları; saplantılı bağlanma özelliği gösterenlerin ise en yüksek düzeyde içselleştirme gösteren grup oldukları gözlenmiştir.

Literatürde romantik kıskançlık kavramı ilişkide her zaman sağlıksız bir duruma işaret etmeyebilir. Örneğin, Attridge (2013) aşk ile simgeleştirilen tepkisel kıskançlığın ilişki için olumlu bir özelliğe sahip olduğunu ifade etmektedir. Şöyle ki, birbirine aşık olan bir çift kıskançlık tepkisi ile ilişkilerine verdikleri değeri ortaya koyabilirler ve bu sayede ilişkilerini daha canlı tutmaları mümkün olabilir. Aynı zamanda, romantik kıskançlığın yaşanmasında veya davranışa dönüşmesinde bireysel, toplumsal ya da kültürel faktörler etkili olabilmektedir. Örneğin, Türkiye'de kıskançlık romantik ilişki içerisinde kabul edilebilir normal bir duygu olarak ele alınmaktadır. Aynı zamanda, "seven insan kıskanır" sözü ile bu duygunun normalleştirildiği söylenebilir ancak ilişkide kıskançlık her zaman sağlıklı bir şekilde yaşanmayabilir ya da davranışa dönüşmeyebilir. Sağlıklı bir şekilde çözülemeyen kıskançlık, ilişkinin olumlu seyrini olumsuza çevirebilir. Şiddetten eş cinayetine kadar farklı şekillerde sonuçlar doğuran romantik kıskançlığın (Pines ve Friedman, 1998) boşanma nedenlerinden biri (Pines, 1992b) olması büyük bir olasılıktır.

Özetle, romantik kıskançlık ilişkilerde daha çok olumsuz yönleri ile ele alınmaktadır. İlişkide sevginin azalması gibi bir süreçten ilişkinin sonlandırılması gibi sonuçlar doğurabilen romantik kıskançlığın sağlıklı bir şekilde ele alınması için üzerinde çalışılması gerekmektedir. Bununla birlikte, Erikson'un gelişim dönemlerinden biri olarak tanımladığı "yakınlığa karşı yalıtılmışlık dönemine" denk gelen üniversite yıllarındaki bireylerin ilişkilerini etkileyebilecek duygusal kıskançlık gibi bir psikolojik yapının araştırılması ayrıca önemlidir. Yakınlığa karşı yalıtılmış sürecinde olan bir birey eğer ilişki içerisinde kendi varlığını koruyarak sevebiliyor ve başkasıyla doyum verici bir paylaşımda bulunabiliyorsa yakınlık duygusunu geliştirebilir buna karşın sevebilme, ilişki kurma veya sürdürme gibi durumları bazı gerekçelerle gerçekleştiremiyorsa yalıtılmışlık yaşayabilir. Birey romantik ilişki içinde olsun ya da olmasın bu duruma ilişkin bir önyargı veya olumsuz düşünce geliştirmişse olası romantik ilişkisini de sağlıklı olmayan ilişkiler üzerine kurabilir.

Duygusal kıskançlığın ilişkiye zarar veren olumsuz özelliği dikkate alındığında, bu duygunun romantik ilişki başlamadan önce fark edilmesi ve ilişkiye zarar vermeyen bir düzeye getirilmesi için

üzerinde çalışılması gereklidir. Bununla birlikte, görece flört şiddetinin arttığı, ilişkilerin daha hızlı şekilde kopabildiği günümüz koşullarında ilişkiye zarar verebilecek nitelikteki olumsuz içerikli ilişki kavramlarından biri olan romantik kıskançlığın ilişkisi olmayan bireylerde ölçülmesi ilişkinin seyri için önemli bir ipucu olabilir. Bir başka deyişle, romantik ilişkisi olmayan üniversite öğrencilerinin ilerde kuracakları romantik bir ilişkide duygusal kıskançlık düzeylerini tespit edebilmek ilişkinin niteliğine yönelik önleyici çalışmaların da temelini oluşturabilir. Özellikle evlilik öncesi psikolojik danışma hizmetleri sunulurken, bu hizmeti sunan merkez veya birimlerde duygusal kıskançlığın ölçülmesi bu konuda yapılan araştırmaların kuramsal ve uygulama alt yapısını da güçlendirebilir.

Araştırmanın Amacı

Bu çalışmanın amacı, Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu'nun geçerlik ve güvenirlik çalışmalarını yapmaktır. Böylece bu araştırmada "Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu geçerli ve güvenilir bir ölçme aracıdır." hipotezi test edilmektedir.

YÖNTEM

Bu araştırma, Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu'nun romantik ilişkisi olmayan üniversite öğrencileri için geçerlik ve güvenirlik çalışmalarının yapıldığı betimsel bir araştırmadır. Betimsel bir araştırmanın, bazı olgu, olay veya durumları açıklamaya çalışan bir araştırma (Boyacı ve Baykuş, 2015) olma özelliğinden hareketle, bu araştırmada üniversite öğrencileri üzerinde duygusal kıskançlık kavramının psikometrik özellikleri açıklanmaya çalışılmıştır.

Çalışma Grubu

Bu araştırmanın çalışma grubunu, 2015-2016 Eğitim-Öğretim Yılı Güz Döneminde Adıyaman Üniversitesi'nde öğrenimlerine devam etmekte olan ve rasgele örnekleme yöntemiyle seçilen toplam 283 dördüncü sınıf öğrencisi oluşturmaktadır.

Birinci çalışma grubu: Bu çalışma grubu, 158'si kadın (%63) ve 92'si erkek (%37) toplam 250 üniversite öğrencisinden oluşmaktadır. Bu öğrencilerin yaş ortalaması 22.41'dir. Bu öğrencilerin 64'ü (%25.6) İslami İlimler Fakültesi'ne, 65'i (%26) Fen-Edebiyat Fakültesi'ne ve 121'i (% 48.4) de Eğitim Fakültesi'ne devam etmektedirler.

İkinci çalışma grubu: Bu çalışma grubu ise test tekrar test analizleri için çalışmaya dahil edilen 18'si kadın (%54,5) ve 15'i erkek (%45.5) toplam 33 üniversite öğrencisiden oluşmaktadır. Bu öğrencilerin yaş ortalaması ise 22.9'dur.

Veri Toplama Araçları

Kişisel bilgi formu

Araştırmacı tarafından geliştirilen Kişisel Bilgi Formu'nda katılımcıların yaş, cinsiyet, öğrenim gördükleri fakülteleri belirlemeye yönelik üç soruya yer verilmiştir.

Eş duygusal kıskançlık ölçeği (EDKÖ)

Evli bireylerin duygusal kıskançlık düzeylerini ölçmek amacıyla Kızıldağ ve Yıldırım (2017) tarafından geliştirilmiştir. Ölçek geliştirme sürecinde ilgili alan yazın incelendikten sonra, romantik ilişkisi olan 10 bireyle görüşme, 33 evli birey üzerinde pilot uygulama, 267 evli birey üzerinde açıklayıcı faktör analizi, 303 evli birey üzerinde ise doğrulayıcı faktör analizi yapılmıştır. Açıklayıcı faktör analizi sonucunda EDKÖ'nün üç faktörlü bir yapıda olduğu görülmüştür. Faktörlerdeki maddeler, içerikleri açısından incelenerek birinci faktöre "değersizlik hissi", ikinci faktöre "ilişkisel doyumsuzluk ve aşkın yitimi" ve üçüncü faktöre ise "beraber zaman geçirmede isteksizlik" isimleri

verilmiştir. EDKÖ'nün doğrulayıcı faktör analizinde ölçekten iyi düzeyde uyum indeksleri elde edilmiştir. Ölçme modelinin genel uyum katsayıları χ^2 204=600.988; p=0.00; x^2/sd = 2.95; GFI = .84; AGFI= .80; NFI= .87; CFI = .91 ve RMSEA= 0.08'dir. EDKÖ'nün güvenirliği iki ayrı veri seti üzerinde test edilmiştir: Birincisi, 267 kişiden elde edilen veri seti üzerinde, EDKÖ'nün Cronbach Alpha katsayısı .951 ve 303 kişiden elde edilen veri üzerinde ise Cronbach Alpha katsayısı .947 bulunmuştur. İkincisi, iki yarı test güvenirliği incelenmiştir. EDKÖ maddeleri tek numaralı ve çift numaralı maddeler olarak ayrılmış ve aralarındaki korelasyon, 267 birey için .925, 303 birey için .922 olarak bulunmuştur. Elde edilen güvenirlik katsayıları EDKÖ'nün evli bireylerin duygusal kıskançlık düzeylerini ölçmek amacıyla güvenle kullanılabileceğini göstermektedir. Ölçek maddeleri "1=Hiç Kıskanmam; 2= Biraz Kıskanırım ve 3=Çok Kıskanırım" şeklinde puanlanmaktadır. Ölçekten alınan yüksek puanlar evli bireylerin ilişkilerindeki duygusal kıskançlık düzeylerinin yüksek olduğu anlamına gelmektedir. Ölçekten alınabilecek puan aralığı 22-66 arasında değişmektedir. Bununla birlikte ölçeğin boyutundaki maddeler istatistiksel olarak "tek puan" vermeye uygundur. İlk boyuttaki maddelerin evlilik ilişkisinde bir eşin diğer eşten hissettiği değersizlik hissi, önem verilmeme gibi olumsuz duyguları içeren maddelerden oluşmasından dolayı bu şekilde adlandırılmıştır. Bu boyuttaki maddelere örnek olarak *"Karşı cinste birilerinin yaptıklarını hoşgörü ile karşılarken aynı durumda benim yaptıklarımı eleştirmesini"* ve *"Beni karşı cinsten birileriyle kıyaslayarak benim beceriksiz, başarısız vb. olduğumu söylemesini"* maddeleri verilebilir. İkinci alt boyuttaki maddelerin evlilik ilişkisinde bir eşin diğer eşin davranışlarından algıladığı ilgisizlik sebebiyle *"ilişkisel doyumsuzluk ve aşkın yitimi"* olarak adlandırılmıştır. Bu boyutta yer alan maddelere *"Karşı cinsten birilerinin dikkatini çekecek davranışlarda bulunması"* ve *"Karşı cinsle ilişkilerine sınır koymamasını"* maddeleri örnek olarak verilebilir. Üçüncü alt boyuttaki maddelerin de evlilik ilişkisinde beraber zaman geçirmedeki isteksizliği ifade ettiğinden *"Beraber zaman geçirmede isteksizlik"* olarak adlandırılmıştır. Bu boyutta yer alan maddelere *"Benimle vakit geçirmek yerine başka işlerle uğraşmayı tercih etmesini"* ve *"Her fırsatta arkadaşlarıyla buluşup görüşmesini"* örnek verilebilir.

Duygu düzenlemede zorluklar ölçeği (DDZÖ)

Duygu düzenlemede yaşanan zorlukları değerlendirmek için Gratz ve Roemer (2004) tarafından geliştirilen ölçeğin Türkçe geçerlik ve güvenirlik çalışması Rugancı ve Gençöz (2010) tarafından gerçekleştirilmiştir. Ölçek 5'li Likert tipinde ve 36 maddeden oluşmaktadır. Ayrıca ölçek netlik, farkındalık, kabul, dürtü, amaç, strateji olmak üzere altı alt boyut içermektedir. Ölçekten alınan yüksek puan ilgili alt boyutlarda daha fazla zorluk yaşandığı anlamına gelmektedir. DDZÖ'nin tüm ölçek iç-tutarlık değerinin .93 alt boyutlar için ise . 80 ile . 89 arasında olduğu belirtilmektedir (Gratz ve Roemer, 2004). Türkçe geçerlik ve güvenirlik çalışmasında da orijinal faktör yapısının elde edildiği ve altı alt boyutun kullanılabileceği belirtilmektedir. Ayrıca geçerlik ve güvenirlik sonuçlarına göre ölçeğin hem genelinin hem de alt boyutlarının geçerli ve güvenir düzeyde olduğu belirtilmektedir.

İlişki ölçekleri anketi (İ.Ö.A.)

Yetişkin bireylerde dört bağlanma prototipini (güvenli, kayıtsız, korkulu ve saplantılı) ölçmeyi amaçlayan İlişki Ölçekleri Anketi Griffin ve Bartholomew (1994) tarafından geliştirmiş ve Sümer ve Güngör (1999) tarafından Türkçe uyarlaması yapılmıştır. Ölçek 7'li Likert tipinde olup 17 maddeden oluşmaktadır. Ölçeğin orijinal formunun alt ölçeklerinin Cronbach alfa değerleri .70 civarındadır. İÖA'nın orijinalinde alt ölçeklerinin Cronbach alfa iç tutarlılık katsayıları 0.41 ile 0.70 arasında bulunmuştur. Ölçeğin Türkçeye uyarlama çalışmasında ise, iç tutarlık katsayılarının 0.27 ile 0.61; test tekrarı sonuçlarının .54 ile .78 arasında değerler aldığı bildirilmiştir.

İşlem

Araştırmanın uygulamaları doğrudan araştırmacı tarafından yapılmıştır. Uygulamalar, öğrencilere araştırmanın amacı açıklanarak ve gönüllü katılımları sağlanarak gerçekleştirilmiştir. Öğrenciler ortalama 15 dakika içinde ölçek setini yanıtlamışlardır.

Verilerin Analizi

Çalışma kapsamında elde edilen veriler, bilgisayar ortamına girildikten sonra gerekli veri ayıklama işlemleri yapılmıştır. Verilerin analizinde SPSS 21 ve AMOS 21 kullanılmıştır. Ölçek setinden eksik dolduran yedi katılımcının ölçekleri çıkarılmış ve analize geriye kalan öğrencilerin verisi dahil edilmiştir.

BULGULAR

Yapı geçerliği

Ölçeğin yapı geçerliğine ilişkin doğrulayıcı faktör analizi, yordama geçerliği çalışmalarına yönelik bulgular aşağıda sırasıyla sunulmuştur:

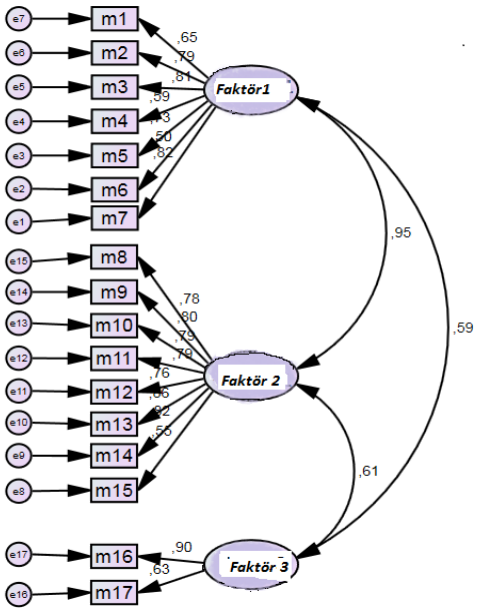
Öncelikle ölçeğin faktörleri arasındaki ilişkilere bakılmış ve bu ilişkilere yönelik korelasyon değerleri Tablo 1'de gösterilmiştir.

Tablo 1. Ölçek Faktörleri Arasındaki Korelasyon Değerleri

Faktörler		1	2	3
1.	Değersizlik Hissi	1		
2.	İlişkisel Doyumsuzluk ve Aşkın Yitimi	.84**	1	
3.	Beraber Zaman Geçirmede İsteksizlik	.51**	.53**	1

***p<0.01*

Tablo 1 incelendiğinde, ölçek faktörleri arasındaki korelasyon değerlerinin .51 ile .84 arasında değiştiği gözlenmektedir *(p<0.01)*. Korelasyon katsayısının mutlak değer olarak, .70-1.00 arasında olması yüksek; 0.70-.30 arasında olması orta ve .30- .00 arasında olması da düşük düzeyde bir ilişkiyi göstermektedir (Büyüköztürk, 2010). Buradan hareketle, bu çalışmadaki ölçek faktörleri arasındaki korelasyon değerlerinin düşük-orta ve yüksek düzeyde ilişkiye sahip olduğu görülmektedir.



Şekil 1. Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu için Doğrulayıcı Faktör Analizi

Şekil 1 incelendiğinde, ölçek faktörlerindeki maddeler arasındaki ilişkiler, değersizlik hissi faktöründe madde faktör yüklerinin (λ) .65 – .82, ilişkisel doyumsuzluk ve aşkın yitimi faktöründe madde faktör yükleri (λ) .55 – .78 aralığında, beraber zaman geçirmede isteksizlik faktöründe ise faktör yükleri (λ) .63 – .90 arasında değiştiği görülmektedir. Doğrulayıcı faktör analizi, daha önceden oluşturulan bir model aracılığıyla gözlenen değişkenlerden yola çıkarak gizil değişken (faktör) oluşturmaya yönelik bir işlemdir (Çokluk, Şekercioğlu ve Büyüköztürk, 2010). Schumacker ve Lomax (2004) doğrulayıcı faktör analizinin en önemli özelliğinin tasarlanan modelin veriye uygun olup olmadığını belirlemek olduğunu ifade eder. Buradan hareketle, üç faktörlü orijinal ölçek yapısının, üniversite öğrencilerinden elde edilen veriye uygun olup olmadığını test edebilmek için sadece doğrulayıcı faktör analizi kullanılmıştır. Bu bilgiye ek olarak, üniversite öğrencileri üzerinde uyarlaması yapılan ölçeğin araştırmacılar tarafından geliştirilen formu üzerinde açıklayıcı faktör analizi yapıldığı için ölçek üzerinde sadece doğrulayıcı faktör analizi yapıldığı söylenebilir. Dolayısıyla, doğrulayıcı faktör analizleri sonucunda üç faktörlü model incelendiğinde χ^2= 334.29, χ^2/df=2.88; RMSEA=.09, NFI=.88, CFI=.92, IFI=.92, TLI=.91 ve RFI=.86 değerleri elde edilmiştir. RMSEA değerinin 0.08 ile 0.10 arasında yer alması orta düzey bir uyuma işaret ederken (MacCallum, Browne ve Sugawara, 1996) ve Kline (2010) χ^2/df ≤3 düzeyinde olmasını kabul edilebilir olarak ifade etmektedir. Aynı zamanda, .00-1.00 arasında değer alan IFI (Artırmalı uyum indeksi), CFI (Karşılaştırmalı uyum indeksi), TLI (Tucker-Lewis indeksi) ve RFI (Göreceli uyum indeksi) değerlerinin 1'e yakın değer alması daha iyi bir uyuma işaret etmektedir (Bentler ve Bonett, 1980). Buradan hareketle, uyum indekslerin iyi bir uyuma işaret ettiği söylenebilir.

Yordama Geçerliği

Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu "değersizlik, ilişkisel doyumsuzluk ve aşkın yitimi ve zaman geçirmede isteksizlik" faktörleri ile Duygu Düzenlemede Zorluklar Ölçeği ve İlişki Ölçekleri Anketi faktörleri ile olan korelasyon değerleri Tablo 2'de verilmiştir.

Tablo 2. DKÖ Üniversite Öğrencileri Formu Faktörlerinin DDZÖ ve İÖA Faktörleri Arasındaki Korelasyon Değerleri

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	,85**	1											
3	,51**	,55**	1										
4	,12	,11	-,01	1									
5	,03	-,01	,06	,35**	1								
6	-,17**	-,09	-,14*	,17*	,05	1							
7	-,20**	-,06	-,11	-,09	-,16*	,47**	1						
8	-,08	-,01	,02	,04	,22**	,44**	,33**	1					
9	-,21**	-,09	-,08	,05	,11	,58**	,57**	,48**	1				
10	,11	,13	-,05	,15*	,07	-,10	-,25**	-,08	-,17**	1			
11	,11	,11	-,05	,06	,07	,06	,07	,12	,14*	,25**	1		
12	-,02	-,02	,04	-,03	,04	,09	,06	,08	,10	-,03	,07	1	
13	-,05	-,02	,02	-,14*	-,07	,18**	,25**	,21**	,22**	-,31**	,28**	,01	1

[***p=0.001; *p=0.05;** 1: edko_değersizlik;2: edko_doyumsuzluk; 3: edko_isteksizlik; 4:ddzö_netlik; 5: ddzö_farkındalık; 6. Ddzö_dürtü; 7:ddzö_kabul; 8: ddzö_amaç; 9: ddzo_strateji; 10: ybbö_güvenli; 11: ybbö_korkulu; 12: ybbö_saplantılı; 13: ybbo_kayıtsız]

Tablo 2 incelendiğinde Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu faktörleri ile DDZÖ ve İÖA ölçek puanları arasındaki korelasyon değerleri incelendiğinde korelasyon değerlerinin -.31 ile ile .85 arasında değiştiği gözlenmektedir. Korelasyon katsayısının mutlak değer olarak, .70-1.00 arasında olması yüksek; 0.70-.30 arasında olması orta ve .30 -0.00 arasında olması da düşük düzeyde bir ilişkiyi göstermektedir (Büyüköztürk, 2010). Buradan hareketle, bu çalışmada ölçek faktörleri arasındaki korelasyon değerlerinin düşük, orta ve yüksek düzeyde ilişkiye sahip olduğu görülmektedir.

Güvenirlik Analizi

Ölçeğin faktörlerine ilişkin yapılan güvenirlik analizleri sonucunda Cronbach alfa değerlerinin .45 ile .91 arasında değiştiği gözlenmektedir. Aynı zamanda iki hafta arayla toplam 33 üniversite öğrencisinden oluşan ikinci çalışma grubu üzerinde yapılan test tekrar test korelasyon katsayısı .97 ($p<0.05$) bulunmuştur. Ölçeğin bütününe ait Cronbach alfa değeri ise .94 bulunmuştur. Güvenirlik katsayının .90 üstünde olması "mükemmel", .80 civarındaki değerlerin "çok iyi" ve .70 civarındaki değerlerin ise yeterli olduğu belirtilmektedir (Kline, 2010). Bu değerler, ölçek faktörlerinin yeterli güvenirliğe sahip olduğunu göstermektedir.

Tablo 3. Duygusal Kıskançlık Ölçeği Üniversite Öğrencileri Formu Madde Toplam Korelasyonları

Faktörler	Maddeler	Düzeltilmiş Madde Toplam Korelasyonları (rjx)
Değersizlik Hissi	m1	.63
	m2	.72
	m3	.75
	m4	.61
	m5	.68
	m6	.52
	m7	.78
İlişkisel Doyumsuzluk ve Aşkın Yitimi	m8	.74
	m9	.74
	m10	.73
	m11	.75
	m12	.73
	m13	.79
	m4	.77
	m15	.59
Beraber Zaman Geçirmede İsteksizlik	m16	.59
	m17	.43

Tablo 3 incelendiğinde ölçeğin madde toplam korelasyonlarının .43 ile .79 arasında değiştiği gözlenmektedir. Korelasyon katsayısının mutlak değer olarak, .70-1.00 arasında olması yüksek; 0.70-.30 arasında olması orta ve .30-0.00 arasında olması da düşük düzeyde bir ilişkiyi göstermektedir (Büyüköztürk, 2010). Buradan hareketle, madde toplam korelasyon değerlerinin orta ve yüksek düzeyde ilişkiye sahip olduğu görülmektedir.

SONUÇLAR VE TARTIŞMA

Duygusal Kıskançlık Ölçeği Üniversite Formu'nun geçerlik ve güvenirlik çalışmalarının yapıldığı bu araştırmada, analizler sonucunda ölçeğin geçerli ve güvenilir bir ölçme aracı olduğu sonucuna ulaşılmıştır. Duygusal Kıskançlığı Ölçeği Üniversite Formu 17 madde ve "değersizlik hissi, ilişkisel doyumsuzluk ve aşkın yitimi ve beraber zaman geçirmede isteksizlik" olmak üzere üç faktörden oluşan bir ölçektir. Değersizlik hissi boyutunda 1, 2, 3, 4, 5, 6 ve 7 numaralı maddeler; ilişkisel doyumsuzluk ve aşkın yitimi boyutunda 8, 9, 10, 11, 12, 13, 14 ve 15 numaralı maddeler ve beraber zaman geçirmede isteksizlik boyutunda ise 16 ve 17 numaralı maddeler yer almaktadır. Faktörlerde yer alan madde sayıları incelendiğinde, üçüncü faktörde sadece iki maddenin olduğu görülmektedir. Raubenheimer (2004) istisna durumlar dışında faktörlerde en az üç maddenin o faktörü daha iyi açıklayabileceğini ifade etmektedir. Bu çalışmada ise 16. ve 17. maddelerin madde toplam korelasyonların kabul edilebilir düzeyde ve maddelerin doğrudan duygusal kıskançlığın literatürdeki kuramsal alt yapısı ile ilişkili olması faktördeki madde sayısının sınırlı fakat bahsedilen nedenlerden ötürü kabul edilebilir olduğunu açıklayabilir. Faktörlerde yer alan madde örneklerine bakıldığında, değersizlik hissi boyutunda "Karşı cinsten birilerinin başarılarını takdir ederken benim başarılarımı küçümsemesini" ve "Hatalar karşısında başkalarını affederken, aynı durumda beni affetmemesini" maddeleri yer almaktadır. İlişkisel doyumsuzluk ve aşkın yitimi olarak isimlendirilen ikinci boyutta "Karşı cinsten birilerinin dikkatini çekecek davranışlarda bulunması" ve "Karşı cinsle ilişkilerine

sınır koymamasını" gibi maddeler bulunurken; beraber zaman geçirmede isteksizlik olarak adlandırılan üçüncü faktörde ise "Benimle vakit geçirmek yerine başka işlerle uğraşmayı tercih etmesini" ve "Her fırsatta arkadaşlarıyla buluşup görüşmesini" maddeleri bulunmaktadır. Buradan hareketle, ölçek maddeleri ve faktörleri literatürde duygusal kıskançlığın kuramsal yapısına (Shackelford ve Buss, 1997) uygun olduğundan bu ölçeğin gerek romantik ilişkisi olan gerekse de romantik ilişkisi olmayan bireylerdeki duygusal kıskançlığın belirlenmesinde işlevsel bir ölçme aracı olduğu düşünülmektedir. Bir başka ifadeyle, duygusal kıskançlık kavramının bireylerde genellikle gözlenebilir bir durum olduğu ve romantik ilişkisi olan veya olmayan bireylerde de var olabilen bir duygu olduğu söylenebilir.

Bununla birlikte, Duygusal Kıskançlığı Ölçeği Üniversite Formu'ndan alınan yüksek puan duygusal kıskançlık düzeyinin yüksekliğine işaret etmektedir. Ölçekten alınabilecek puan aralığı 17-51 arasında değişmektedir. Üçlü Likert tipinde olan ölçekte maddeler "1=Hiç Kıskanmam; 2= Biraz Kıskanırım ve 3=Çok Kıskanırım" şeklinde puanlanmaktadır. Ölçek maddeleri arasında reverse (tersten puanlanan) madde olmadığından sadece düz puanlama yapılmaktadır. Ölçeğin geçerlik analizleri kapsamında yapı geçerliği ve yordama geçerliğine bakılmıştır. Doğrulayıcı faktör analizleri sonucunda ölçeğin orijinalindeki üç faktörlü yapısı doğrulanmıştır. Buradan hareketle, ölçeğin gerek evli bireyler üzerinde geliştirilmiş hali gerekse de romantik ilişkisi olmayan üniversite öğrencileri üzerindeki yapılan hali ile duygusal kıskançlık ile ilgili benzer yapıları ortaya koyduğu söylenebilir. Aynı zamanda analiz sonucunda model uyum indeksleri; χ^2= 334.29, χ^2/df=2.88; RMSEA=.09, NFI=.88, CFI=.92, IFI=.92, TLI=.91 ve RFI=.86 bulunmuştur. RMSEA değerinin 0.08 ile 0.10 arasında yer alması orta düzey bir uyuma işaret ederken (MacCallum, Browne ve Sugawara, 1996) ve Kline (2010) χ^2/df ≤3 düzeyinde olmasını kabul edilebilir olarak ifade etmektedir. Aynı zamanda, .00-1.00 arasında değer alan IFI (Artırmalı uyum indeksi), CFI (Karşılaştırmalı uyum indeksi), TLI (Tucker-Lewis indeksi) ve RFI (Göreceli uyum indeksi) değerlerinin 1'e yakın değer alması daha iyi bir uyuma işaret etmektedir (Bentler ve Bonett, 1980). Yordama geçerliği kapsamında, Duygusal Kıskançlık Ölçeği Üniversite Formu'nun "değersizlik, ilişkisel doyumsuzluk ve aşkın yitimi ve beraber zaman geçirmede isteksizlik" faktörleri ile Duygu Düzenlemede Zorluklar Ölçeği ve İlişki Ölçekleri Anketi faktörleri arasındaki korelasyon değerleri incelenmiştir. Yapılan korelasyon analizi sonucunda, değerlerin -.31 ile ile .85 arasında değiştiği gözlenmiştir. Bununla birlikte, güvenirlik analizleri kapsamında Cronbach alfa güvenirlik katsayısına, madde toplam korelasyonlarına bakılmış ve test-tekrar test güvenirliği incelenmiştir. Ölçeğin Cronbach alfa değeri .94; test-tekrar test korelasyon katsayısı .97 bulunmuştur. Güvenirlik katsayısının, $0.00 \leq \alpha \leq 0.40$ ise ölçek maddeleri güvenilir değil; $0.40 \leq \alpha \leq 0.60$ ise ölçeğin güvenirliği düşük; $0.60 \leq \alpha \leq 0.80$ olması ise ölçek oldukça güvenilir ve $0.80 \leq \alpha \leq 1.00$ ise ölçek yüksek düzeyde güvenilir bir ölçek (Kalaycı, 2010) olduğunu göstermesi bu çalışma kapsamında ele alınan Duygusal Kıskançlık Ölçeği Üniversite Formu'nun da iyi bir psikometrik bir özelliğe sahip olduğunu göstermektedir. Aynı zamanda, ölçek madde toplam korelasyon değerlerinin .43 ile .79 arasında değiştiği gözlenmektedir. Korelasyon katsayısının mutlak değer olarak, .70-1.00 arasında olması yüksek; 0.70-.30 arasında olması orta ve .30-0.00 arasında olması da düşük düzeyde bir ilişkiyi göstermektedir (Büyüköztürk, 2010).

ÖNERİLER

Duygusal Kıskançlık Ölçeği Üniversite Formu'nun geçerlik ve güvenirliğine ilişkin elde edilen kanıtlar, ölçeğin romantik ilişkisi olmayan üniversite öğrencilerinin duygusal kıskançlık düzeylerini geçerli ve güvenilir olarak ölçecek nitelikte olduğunu göstermektedir. Buradan hareketle, romantik ilişkiler üzerinde çalışan araştırmacı ve uygulamacılar, Duygusal Kıskançlık Ölçeği Üniversite Formu'nu üniversite öğrencileri üzerinde yapacakları çalışmalarda güvenle kullanılabilirler. Bu sayede romantik ilişkisi olmayan üniversite öğrencilerinin gelecekte kuracakları romantik ilişkilerine zarar verme olasılığı olan duygusal kıskançlık durumlarını önceden tespit etmeleri mümkün olabileceğinden önleyici çalışmalar için de temel oluşturabilir. Aynı zamanda, Evlilik Öncesi Psikolojik Danışma alanında çalışan psikolojik danışmanlar özellikle deneysel çalışmalarda duygusal kıskançlık durumunun zararlı etkilerini azaltmaya yönelik çalışmaları da mümkün olabilir. Ayrıca,

ilerde yapılacak betimsel çalışmalarda duygusal kıskançlık, bağlanma ve duygu düzenlemede zorluklar arasındaki ilişkiler incelenebileceği gibi kültürel bağlamda duygusal kıskançlığın ilişkilerin niteliğine etkileri de araştırılabilir.

KAYNAKÇA

Attridge, M. (2013). Jealouys and relationship closeness: Exploring the good (reactive) and ban (suspicious) sides of romantic jealousy. *Sage Open, 3*(1), 1-16.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Boyacı, A. ve Bozkuş, K. (2015). *Araştırma yöntemleri desen ve analiz* (Çev. A. Aypay). Ankara: Anı Yayıncılık.

Buunk, B. (1997). Personality, birth order, and attachment styles as related to various types of jealousy. *Personality and Individual Differences, 23*(6), 997–1006.

Büyüköztürk, Ş. (2010). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Akademi.

Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2010). *Sosyal bilimler için çok değişkenli istatistik.* Ankara: Pegem Akademi.

Demirtaş-Madran, A. (2008). Duygusal ve cinsel kıskançlık açısından temel cinsiyet farklılıkları: Evrimsel yaklaşım ve süregelen tartışmalar. *Türk Psikiyatri Dergisi, 19*(3), 300-309.

Edalati, A., & Redzuan, M. (2010). The relationship between jealousy and agression: A review of literatures related to wives' aggression. *European Journal of Scientific Research, 39*(4), 498-504.

Gentzler, A. L., Kerns, K. A., & Keener, E. (2010). Emotional reactions and regulatory responses to negative and positive events: Associations with attachment and gender. *Motivation and Emotion, 34*(1), 78-92.

Gratz, K. L., Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation. *Journal of Psychopathology & Behavioral Assessment, 26*, 41-54.

Griffin, D. W., & Bartholomew, K. (1994). Models of the self and other: Fundamental dimensions underlying measures of adult attachment. *Journal of Personality and Social Psychology, 67*(3), 430-445.

Karakurt, G. (2001). *Yetişkin bağlanma stillerinin romantik kıskançlık üzerindeki etkileri (*Yüksek lisans tezi, Orta Doğu Teknik Üniversitesi, Psikoloji Anabilim Dalı, Ankara). http://tez2.yok.gov.tr/ adresinden edinilmiştir.

Kızıldağ, S., Yıldırım, İ. (2017). Eş duygusal kıskançlık ölçeği'nin geliştirilmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 17*(1), 175-190.

Kline, R. B. (2010). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Knobloch, L. K., Solomon, D. H., & Cruz, M. G. (2001). The role of relationship development and attachment in the experience of romantic jealousy. *Personal Relationships, 8*(2), 205–222.

Levy, K. N., & Kelly, K. M. (2010). Sex differences in jealousy: A contribution from attachment theory. *Psychological Science, 21*(2), 168-173.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

Pfeiffer, S. M., & Wong, P. T. P. (1987). Multidimensional jealousy. *Journal of Social and Personal Relationships, 6*, 181–196.

Pines, A. M. (1992a). Romantic jealousy: Five perspectives and an integrative approach. *Psychotherapy, 29*(4), 675-683.

Pines, A. M. (1992b). *Romantic jealousy: Understanding and conquering the shadow of love.* New York: St. Martin's.

Pines, A. M., & Friedman, A. (1998). Gender differences in romantic jealousy. *The Journal of Social Psychology, 138*(1), 54–71.

Raubenheimer, J. (2004). An item selection procedure to maximise scale reliability and validity. *Journal of Industrial Psychology, 30*(4), 59-64.

Rugancı, R. N., & Gençöz, T. (2010). Psychometric properties of the Turkish version of the difficulties in emotional regulation scale. *Journal of Clinical Psychology, 66*, 442 – 455.

Schumacker, R. E., & Lomax, R. G. (2004). *A Beginner's Guide to Structural Equation Modeling.* New Jersey: Taylor & Francis.

Selterman, D. F., & Maier, M. A. (2013). Secure attachment and material reward both attenuate romantic jealousy. *Motive Emotion, 37*, 765–775.

Shackelford, T. K., & Buss, D. M. (1997). Cues to infidelity. *Peronality and Social Psychology Bulletin, 23*, 1034–1045.

Sharpsteen, D. J., Kirkpatrick, L. A. (1997). Romantic jealousy and adult romantic attachment. *Journal of Personality and Social Psychology, 72*(3), 627-640.

Simpson, J. A. (1990). Influence of attachment styles on romantic relationships. *Journal of Personality and Social Psychology, 59*(5), 971–980.

Sümer, N. ve Güngör, D. (1999). Yetişkin bağlanma stilleri ölçeklerinin Türk örneklemi üzerinde psikometrik değerlendirmesi ve kültürlerarası bir karşılaştırma. *Türk Psikoloji Dergisi, 14*(43), 71-106.

White, G. L. (1981). Some correlates of romantic jealousy. *Journal of Personality, 49(2),* 129-147.

Wreen, M. J. (1989). Jealousy. *Nous, 23*(5), 635-652.

EXTENDED ABSTRACT

Introduction

Romantic jealousy is defined as a whole of complex emotions, thoughts, and behaviors, which include threats against individual's self-worth or relationship (White, 1981). The common behavior that can be observed in both sexual and emotional jealousy, which are classified as two forms of romantic jealousy, is one of the spouse making less effort for the relationship and the reluctance for being attractive and desirable for the the spouse (Shackelford and Buss, 1997). Behaviors that unveil emotional jealousy are sorted by Shackelford and Buss (1997) as relational dissatisfaction and loss of love; emotional neglect; reluctance for spending time together; passive rejection and starting to behave inconsiderately; communicating in an angry, critical, and interrogative way; avoiding from talking about a specific person; and adopting a guilty and nervous communication style. In the literature, romantic jealousy in relationships may generally be associated with negative emotions. While Pfeiffer and Wong (1989) explain emotional jealousy with some feelings like fear, anger, lack of confidence, sadness; Wreen (1989) states that jealousy grounds on lack of confidence, fear of losing, and emotional alienation. Under current circumstances with increasing divorce rates and faster relationship breakdowns, it is a preventive action to evaluate jealousy, which is one of the relationship concepts with a negative content that may damage the relationship, also in individuals, who are not involved in a romantic relationship. In other words, to be able to determine the emotional jealousy level of the university students, who are not currently involved in a romantic relationship, in their future romantic relationships can underlie preventive studies regarding the quality of relationship. From this point of view, the aim of this study is to make validity and reliability studies of Spouse Emotional Jealousy Scale (Kızıldağ and Yıldırım, 2017), which was developed on married individuals, on university students, who are not involved in a romantic relationship. Thereby, this study seeks an answer to the question of "Is the University Form of Emotional Jealousy Scale a valid and reliable measurement tool?

Method

University Student Form of Emotional Jealousy Scale is a descriptive research, within which validity and reliability studies for university students, who are not involved in a romantic relationship, are conducted. There are two study groups in this research. The first study group consisted of a total of 250 university students with 158 students were female (63%) and 92 students were male (37%) and the second study group consisted of a total of 33 university students with 18 students were female (54.5%) and 15 students were male (45.5%), who were included in the study for test-retest reliability analyses. In the study, Personal Information Form consists of three questions to determine the age, gender, and faculty of the participants was used. Spousal Emotional Jealousy Scale that was developed by Kızıldağ and Yıldırım (2017) with the aim of measuring emotional jealousy level of married couples was used for validity and reliability studies. For predictive validity, Difficulties in Emotion Regulation Scale, which was developed by Gratz and Roemer (2004) and of which Turkish validity and reliability studies were conducted by Rugancı and Gençöz (2010), was used. Developed by Griffin and Bartholomew (1994) and adopted to Turkish by Sümer and Güngör (1999) Relation Scales Questionnaire, which aims to measure four attachment prototypes (secure, dismissive avoidant, fearful, preoccupied) in adults, was also used. For data analysis, SPSS 21 and AMOS 21 programs were used.

Results and Discussion

In the scope of validity analysis, construct validity and predictive validity were examined. Cronbach alpha reliability coefficient, total-item correlation, and test-retest reliability were examined in the scope of reliability analysis. Model fit indices were found as χ^2= 334.29, χ^2/df=2.88; RMSEA=.09, NFI=.88, CFI=.92, IFI=.92, TLI=.91, and RFI=.86 as a result of confirmatory factor analysis. A RMSEA value between 0.08 to 0.10 indicates a medium level cohesion (MacCallum, Browne ve Sugawara, 1996). Kline (2010) also claimed that the model fit is acceptable, if χ^2/df ≤3. As a result of the correlation analysis, it is observed that the values change between -.31 and .85. In addition to this, Cronbach alpha reliability coefficient, total-item correlation, and test-retest reliability were examined in the scope of reliability analysis. As a result of the reliability analysis, which was carried out concerning the factors of the scale, it is observed that Cronbach alpha values change between .45 and .91. Test-retest correlation coefficient is .97 *(p<0.05).*

Cronbach alpha value that belongs to the whole scale was found as .94. Meanwhile, it is observed that the item-total correlation values of Emotional Jealousy Scale University Form change between .43 and .79. The evidence on the validity and reliability of Emotional Jealousy Scale University Form shows that the scale is capable of measuring the emotional jealousy status of the university students, who are not involved in a romantic relationship, in a valid and reliable way. Emotional Jealousy Scale University Form is a scale that consists of 17 items and three factors that are "feeling of worthlessness, relational dissatisfaction and loss of love and unwillingness for spending time together". A high score on the scale indicates a high level of jealousy. The scores range between 17 and 51. The ratings on this three-point Likert-type scale are "1= I am never jealous; 2=I am a little jealous; 3=I am very jealous". Since there are no reverse items among the scale items, straight grading is made. To sum up, it can also provide a basis for preventive studies, as it is possible to identify the jealousy statuses of the university students that might damage their future romantic relationships, through the scale that can measure the emotional jealousy levels of the university students, who are not involved in a romantic relationship, in a valid and reliable way. At the same time, this scale can be used by the counselors that are working in the Premarital Counseling, to determine the jealousy statuses especially in experimental studies and thus studies can be conducted to decrease the harmful effects of jealousy.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2017; 8(1);158-168

Interview with Stephen G. Sireci on Validity

Interviewer: Nuri DOĞAN*

Interviewee: Stephen G. SIRECI**

INTRODUCTION

The following questions were created within the scope of Classical Test Theory course (OLC720) offered in the doctoral programme of Measurement and Evaluation in Education branch in the Educational Sciences Department of Hacettepe University. The question of "what three questions about validity would you ask if an internationally famous expert was in front of you?" was asked 23 students taking the course, and the participants were asked to write down their questions. 69 questions in total were revised by the lecturer of the course in terms of scope, importance and clarity, and the number of questions was reduced to 37 by removing 32 of them. The students participating in the classes were then asked to order the questions selected by the lecture from the most important to the least important. The order of importance for each question was determined by adding up the scores given by the students. 15 out of the 37 questions which had been ordered according to the students' rating were selected to be answered.

QUESTIONS and ANSWERS of INTERVIEW

1. ***What does the fact that different types of evidence have increased and that there are no criteria as to what type of evidence we should prioritize make you think about classifications in relation to types of validity? What is the most valid definition and classification of validity in your opinion?***

2. ***There are different approaches in defining validity types. For example, some methods of gathering evidences for validity are mentioned in the last version of Standards Book, but not validity types as content, criterion-related or construct validity. What are the reasons for disagreements concerning these different approaches?***

I will answer these two questions together. I think the best definition of validity is provided by the current version of the *Standards for Educational and Psychological Testing*. The current version was published in 2014, and it is the 6th version. They define validity as, "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014, p. 11). This definition is essentially the same as that provided in the 5th edition (AERA, APA, & NCME, 1999). It is important because from the definition we can see that,

- Validity must be evaluated with respect to a particular *purpose* or *use* of a test.
- Tests are not "inherently" valid or invalid. What must be validated (that is, supported by research and theory) is the use of a test for a particular purpose.

*Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey e-mail:nurid@hacettepe.edu.tr
**Prof. Dr. ,University of Massachusetts Amherst , College of Education, Amherst–United States, e-mail: sireci@acad.umass.edu

- Validation requires both evidence and theory to support the use of a test for a particular purpose.

There is not one study that can be done to validate the use of a test for a specific purpose. There are different types of validity evidence, and the types of evidence used to defend the use of a test for a particular purpose will vary based on the purpose of the test. The last two versions of the *Standards* specify five sources of evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (AERA et al., 2014, p. 13). These sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing.

These sources of validity evidence described by the *Standards* (AERA et al., 1999, 2014) are not the same as described in earlier versions of the *Standards*. In Table 1, I list the different versions of the *Standards* and how they described different categories of validity or types/sources of validity evidence. Note that the current version describes "sources of validity evidence" because there are not different types of validity. Validity is a unitary concept, whether you put the word "construct" in front of it, or not.

I think the current (AERA et al. 2014) definition of validity, and the five sources of validity evidence are the best ways to describe validity for several reasons. First, they are not the product of one person. For over 60 years, three organizations have worked together to come to some consensus about what validity means and how test scores (uses) should be validated. Second, the definition emphasizes that validity refers to test use, and that validation requires both theoretical justification and empirical evidence. These are truisms that are hard to reject.

Tablo 1. Categorization of Validity Evidence Over Time in the *Standards*

Publication	Validity Classifications
Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal (APA, 1952)	Categories: predictive, status, content, congruent
Technical recommendations for psychological tests and diagnostic techniques (APA, 1954)	Types: construct, concurrent, predictive, content
Standards for educational and psychological tests and manuals (APA, 1966)	Types: criterion-related, construct-related, content-related
Standards for educational and psychological tests (APA, AERA, & NCME, 1974)	Aspects: criterion-related, construct-related, content-related
Standards for educational and psychological testing (AERA, APA, & NCME, 1985)	Categories: criterion-related, construct-related, content-related
Standards for educational and psychological testing (AERA, APA, & NCME, 1999)	Sources of evidence: content, response processes, internal structure, relations to other variables, consequences of testing

3. *How is the response process a test used as evidence for validity? Can response process be used as evidence for validity in examinations with extensive participation?*

Validity evidence based on response processes refers to "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers" (AERA et al., 2014, p. 15). Examples of this type of evidence include interviewing examinees about their responses to test questions, systematic observations of examinees responding to test items, evaluation of the criteria used by judges when scoring performance tasks, analysis of item response time (chronometric analysis), tracking students' eye movements, and evaluation of the reasoning processes examinees use when solving test items. This evidence is particularly useful for evaluating the degree to which tests tap higher-order skills and for evaluating how well students in different subpopulations understand the test items. Given that many new educational tests emphasize higher-level cognitive skills, evidence will be needed that these tests adequately measure these skills.

4. *Should different ways be followed in collecting validity evidence in cases when absolute evaluation and relative evaluation are made?*

I don't understand what absolute and relative evaluation mean. However, let's talk about how different sources of validity evidence are put together to make a "validity argument." Kane (1992, 2006, 2013), suggested that validating the use of a test for a particular purpose is tantamount to developing a sound and logical argument that use of the test for a particular purpose is justified. The *Standards* essentially adopted this perspective by claiming that the five sources of evidence should be coherently synthesized to support use of a test for a particular purpose. For example, they state "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA et al., 2014, p. 21).

The argument-based approach to validation is similar to defending the use of a test in a courtroom. The idea is to present a preponderance of evidence that would support the use of a test for a particular purpose. This body of evidence should include evidence that the test is fulfilling its intended objectives and is not producing undesired consequences.

5. *What can be done to collect evidence for validity if adequate sample is not available to perform validity study?*

Validity evidence based on test content is typically gathered using subject matter experts, and usually represent very small numbers of experts (e.g., 10 or fewer). So, please remember validity studies are not all based on analysis of item responses or test scores. Content validity evidence is fundamental and necessary for educational tests. Much of the research on validity evidence based on response processes also uses very small sample sizes.

6. *How can the response process of a test be used as evidence of validity? Is it possible to use the response process as evidence for validity in exams with large rates of participation?*

See answer to question #3.

7. *How do we explain the relations between multiple methods and multiple traits analysis?*

I recommend reading Campbell and Fiske (1955).

8. *Boud (1995) and Messick (1995), have raised the concept of consequential validity for alternative assessment methods. Which techniques are used in determining the consequential validity based on the influence of the assessment on learning?*

Messick did not use the term "consequential validity." The *Standards* describe "validity evidence based on consequences of testing." Validity evidence based on consequences of testing refers to evaluating the intended and unintended consequences associated with a testing program. Tests are used to promote positive consequences such as appropriate diagnosis of psychological disorders, protection of the public, improved instruction, and better understanding of the constructs measured. Unintended positive consequences that were not explicitly intended or envisioned may also emerge. However, unintended negative consequences may also occur in a testing program. Examples of unintended consequences may be adverse impact that leads to decreased education and employment opportunities for certain groups, increased dropout rates in schools, and poor decisions regarding resource allocations or employees' salaries based on test performance.

Validity evidence based on the consequences is particularly important in considering the validation of tests for some purposes, such as accountability (e.g., using tests to evaluate schools or teachers). In the USA, accountability testing is required by federal educational policy, and typically comes with a

theory of action outlining the intended consequences for stakeholders. For example, using students' test results to evaluate teachers encourages teachers to teach the intended curriculum, and it is assumed this more focused instruction will improve student learning with respect to that curriculum. The degree to which these intended consequences are realized, and other, unintended consequences (e.g., decreased teacher morale, narrowing the curriculum in a way that decreases student learning) are minimized is essential to investigating the validity of educational tests for accountability purposes. Other testing purposes, such as using a test for high school graduation or college admission, also have consequences that should be evaluated.

9. ***There are also different approaches and classifications on the validity in the literature. The question is based on a hypothetical situation in order to privatize the situation. -You want to measure top-level skills (e.g. problem solving, critical thinking). The result of test scores will be used in decisions that have a high stakes qualification. How do you provide validity evidence in the process of developing the test to the meaning and use of the scores to ensure validity?***

See answer to #4.

10. ***Scores for individuals are calculated by adding up the numerical values of responses to items in Likert type scales. Yet, it is also clear that the degree to which each item serves to the relevant structure differs. How correct is it to collect the scores in a straightforward way and how high is the validity of the findings/inferences made on the basis of those scores?***

Remember that validity refers to the use of a test for a particular purpose. There are typically no one-item tests and so inferences and decisions are made on the basis of scores calculated across many items. If there is a problem with an item or two, it may or may not be important, depending on how it affects the total test score and the interpretation. It is good for validity analysis to include analyses at different levels, such as the item level, total score level, and subscore level. If you consider the 5 sources of validity evidence, all levels are accounted for. Item analyses and differential item functioning are part of validity evidence based on internal structure. Dimensionality analyses also typically focus on the item level, as do studies based on test content. However, relations with other variables (e.g., predictive validity, differential predictive validity, MTMM) will focus at the total test score level. The specific validity question to be evaluated will dictate the level of analysis.

11. ***We rather focus on test reliability in Turkey and we see that our institutions have not put most of the applications concerning validity into practice yet. Let us assume that a commission authorised in tests in Turkey would like to work with you. What applications and how would you like to change by considering validity on the basis of scientific and social values? What would you recommend?***

Reliability is important, but it is more important to remember that tests that are not valid for a purpose can still be reliable. For example, a college admissions test may produce reliable scores, but it would not be valid for assigning grades to students in a math course. I think my answers to the previous questions describe how I would go about the process of test validation.

12. ***It is known that visually impaired students are given test booklets printed in differing font sizes according to the degree of their sight and students who cannot see are not held responsible for visual questions and are asked to answer fewer questions with extra time with the support of a reader in examinations held by Centre for Measurement- Selection and placement in Turkey. In addition to that, there are differing practices for those who certify their handicap. By taking the above mentioned situations into consideration, should validity evidence for a test be considered on normal conditions, or should different validity evidence be searched for the sub-groups of different characteristics?***

It depends what the validity questions are. Remembering that validity refers to a specific testing purpose, one question might be if the test is similarly valid for standard test administrations and accommodated test administrations. Another question might be if the scores from standard and accommodated administrations are comparable. A more specific question might be whether the accommodation has changed the construct measured. There is a great deal of literature and applied research in these areas.

13. ***Can a solution be found to the problem of range narrowing with a statistical approach?***

Restriction of range is really important when evaluating test data. I have seen many studies that have concluded a lack of invariance across groups (at either the item or total test score level), that is probably just picking up on differential restriction of range. Statistics such as item biserials, reliability coefficients, factor loadings, and etc., need variability. If a test is too easy or too hard for one group, it will look like a source of bias, but really it is just an artifact of restriction of range.

One way we have handled this problem is to sample from the unrestricted group in a way that matches the distribution in the restricted group. That strategy controls for differential restriction of range. Of course, disattenuation for restriction of range is also handy when appropriate, and when you have the data to do it.

14. ***Messick (1995) states that validity can be defined broadly as the result and use of both evidence collection and score interpretation and sees content validity and criterion-based validity as sub-parts of construct validity. Messick points out that the concept of unitary validity demonstrates the construct validity of a test. The concept was criticised in that it could not answer simple questions about what a test measured and that it was rather related with complex score interpretations which were explained with nomological network. Considering the fact that different types of validity serve to different purposes, is it possible to form an umbrella term for validity (as different from construct validity)?***

15. ***Messick (1995) states that validity can broadly be defined as the result and use of both evidence collection and score interpretation and sees content validity and criterion-based validity as sub-parts of construct validity. Accordingly, how correct is it to investigate construct validity independently of content and criterion-based validity? Should the validity of a measurement always be investigated as a whole? Which type of validity should firstly be looked at?***

Please see the Sireci (2012) for an answer to these two questions.

Stephen G. Sireci ile Geçerlik Üzerine Söyleşi

Söyleşiyi yapan: Nuri DOĞAN[1]
Söyleşi yapılan: Stephen G. SIRECI[2]

Çeviri: Nuri DOĞAN
Burcu ATAR[3]
Nermin KIBRISLIOĞLU UYSAL[4]
Osman TAT[5]
Cem MALAKÇIOĞLU[6]

GİRİŞ

Aşağıdaki sorular Hacettepe Üniversitesi, Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı doktora programındaki OLC720 Klasik Test Kuramı dersi kapsamında üretilmiştir. Derse katılan 23 öğrenciye "Eğer karşınızda 'geçerlik' konusunda uluslararası tanınırlığa (üne) sahip bir uzman olsaydı, ona geçerlikle ilgili sormak istediğiniz en önemli üç soru ne olurdu?" diye sorulmuş ve bu soruları yazmaları istenmiştir. Böylece 69 soru elde edilmiştir. Elde edilen 69 soru kapsam, önemlilik ve anlaşılabilirlik bakımından ders sorumlusu tarafından gözden geçirilmiş ve 32 soru çıkarılarak 37 soruya indirgenmiştir. Derse katılan öğrencilerden ders sorumlusu tarafından seçilen soruları en önemliden en önemsize doğru sıralamaları istenmiştir. Öğrencilerin verdiği puanlar toplanarak soruların önem sırası belirlenmiştir. Öğrencilerin puanlarına göre sıralanmış 37 sorudan 15'i cevaplandırılması için seçilmiştir.

Cevapların çevirisi beş farklı araştırmacı tarafından bağımsız olarak yapılmış sonrasında aralarındaki uyuma bakılmıştır. Çeviri sırasında bire bir çeviri yerine anlam bütünlüğü en doğru ifade eden cümleler kullanılmıştır.

SÖYLEŞİ SORULARI ve YANITLARI

1. ***Farklı kanıt türlerinin çoğalması ve farklı kanıt türlerinden hangisine öncelik vereceğimize ilişkin bir ölçütün olmaması geçerlik türlerine ilişkin yapılan sınıflandırmaların gerekliliği noktasında size ne düşündürtmektedir? Bu kapsamda, size göre en geçerli geçerlik tanımı ve sınıflandırması nedir?***

2. ***Geçerlik türlerinin tanımlamada farklı anlayışlar bulunmaktadır. Örneğin Standartların son versiyonunda geçerlik kanıtları elde etme yolları vardır. Kapsam, ölçüt dayanaklı, yapı geçerliği gibi türler yoktur. Bu ilgili tanımlamalara yönelik görüş ayrılığı ne gibi nedenlerden dolayı ortaya çıkmıştır?***

1-2. Bu iki soruyu birlikte cevaplayacağım. Geçerliğin en iyi tanımının Eğitimsel ve Psikolojik Testler için Standartların güncel baskısında verildiğini düşünüyorum. Güncel baskı 2014 yılında yayımlandı ve bu 6.baskıdır. Onlar geçerliği "kanıt ve kuramın testlerin amaçlanan kullanımları için test

[1] Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta:nurid@hacettepe.edu.tr
[2] Prof. Dr., University of Massachusetts Amherst , Amherst –Birleşik Devletler, e-posta: sireci@acad.umass.edu
3 Doç. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta: burcua@hacettepe.edu.tr
[4] Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta: nkibrislioglu@hacettepe.edu.tr
[5] Arş. Gör., Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye e-posta: osman.tat@hacettepe.edu.tr
[6] Arş. Gör., İstanbul Medeniyet Üniversitesi, Edebiyat Fakültesi, İstanbul-Türkiye e-posta: cemm@medeniyet.edu.tr

puanlarının yorumlanmasını ne ölçüde desteklediği" şeklinde tanımlamaktadır (Amerikan Eğitimsel Araştırma Derneği, Amerikan Psikoloji Derneği ve Eğitimde Ölçme Ulusal Konseyi, 2014, p. 11). Bu tanım esasen 5. baskıda verilen tanımla aynıdır (AERA, APA, & NCME, 1999). Bu tanım önemlidir çünkü tanımda şunları görebiliriz;

- Geçerlik bir testin belli bir amacına veya kullanımına göre değerlendirilmelidir.
- Testler "doğası gereği" geçerli veya geçersiz değildir. Geçerlenmesi gereken (tabii ki, kuram ve uygulama ile desteklenerek) testin belli bir amaç için kullanımıdır.
- Geçerleme hem kanıt hem de kuramın bir testin belli bir amaç için kullanımını desteklemesini gerektirir.

Bir testin belirli bir amaçla kullanımını geçerlemek (validate) için yapılabilecek tek bir çalışma yoktur.

Geçerlik kanıtının farklı türleri vardır ve bir testin belli bir amaç için kullanımını desteklemek üzere kullanılan kanıt türleri testin amacına göre değişecektir. Standartların son iki baskısı "belli bir kullanım için test puanlarının amaçlanan yorumlanmasının geçerliğinin değerlendirilmesinde kullanılabilecek" (AERA ve diğerleri, 2014, p. 13) kanıtların beş kaynağını belirtir. Bu kaynaklar (a) test içeriğine, (b) yanıt süreçlerine, (c) içyapıya, (d) diğer değişkenlerle ilişkilere ve (e) testin sonuçlarına dayanan geçerlik kanıtlarıdır.

Standartların son iki baskısında (AERA ve diğerleri, 1999, 2014) tanımlanan geçerlik kanıtlarının bu kaynakları Standartlar'ın daha önceki baskılarında tanımlananlar ile aynı değildir.

Tablo 1'de Standartlar'ın farklı baskılarını ve geçerliğin farklı kategorilerini veya geçerlik kanıtı türlerini/kaynaklarını nasıl tanımladıklarını listeliyorum. Güncel baskının "geçerlik kanıt kaynaklarını" tanımladığına dikkat edin, çünkü geçerliğin farklı türleri yoktur. Önüne "yapı" kelimesini koysanız da koymasanız da geçerlik bütünsel bir kavramdır. Geçerliğin güncel tanımının (AERA ve diğerleri, 2014) ve geçerlik kanıtlarının beş kaynağının birçok nedenden dolayı geçerliği tanımlamanın en iyi yolu olduğunu düşünüyorum. İlk olarak, bunlar tek bir kişinin ürünü değildir. 60 yılı aşkın bir süredir, üç kuruluş geçerliğin ne anlama geldiği ve test puanlarının (kullanımlarının) nasıl geçerli kılınması gerektiği hakkında fikir birliğine varmak için birlikte çalışmaktadır. İkinci olarak, tanım geçerliğin, test kullanımına işaret ettiğini vurgulamaktadır ve bu geçerleme hem kuramsal doğrulama hem de deneysel kanıt gerektirmektedir. Bunlar reddetmesi zor, apaçık gerçeklerdir.

Tablo 1. Standartlar 'da Geçerlik Kanıtlarının Zaman İçindeki Sınıflandırılması

Yayım	Geçerlik Sınıflandırmaları
Psikolojik testler ve tanılayıcı yöntemler için teknik tavsiyeler: Bir ön öneri (APA, 1952)	Sınıflamalar: Yordama, durum, kapsam, Uyum (uygunluk)
Psikolojik testler ve tanılayıcı yöntemler için teknik tavsiyeler (APA, 1954)	Türler: Yapı, eşzamanlı, yordama, kapsam
Eğitimsel ve psikolojik testler ve kılavuzlar için standartlar (APA, 1966)	Türler: Ölçüt dayanaklı, yapı dayanaklı, kapsam dayanaklı
Eğitimsel ve psikolojik testler için standartlar (APA, AERA, ve NCME, 1974)	Yönler: Ölçüt dayanaklı, yapı dayanaklı, kapsam dayanaklı
Eğitimsel ve psikolojik testler için standartlar (AERA, APA, ve NCME, 1985)	Sınıflamalar: Ölçüt dayanaklı, yapı dayanaklı, kapsam dayanaklı
Eğitimsel ve psikolojik testler için standartlar (AERA, APA, ve NCME, 1999)	Kanıt Kaynakları: Kapsam, yanıtlama süreci, iç yapı, diğer değişkenlerle ilişkiler, testin sonuçları

3. Bir testin cevaplanma süreci geçerlik kanıtı olarak nasıl kullanılmaktadır? Geniş katılımlı sınavlarda cevaplama süreci geçerlik için kanıt olarak kullanılabilir mi?

Yanıtlama sürecine dayalı geçerlik kanıtları "yapı ve testi alan alanlar tarafından gerçekten ortaya konan performans ya da tepkinin ayrıntılı doğası arasındaki uyumla ilişkili kanıtlara" karşılık gelmektedir. (AERA ve diğerleri, 2014, p. 15). Bu kanıt türünün örnekleri katılımcılarla test sorularına verdikleri yanıtlar hakkında görüşme yapmayı, test maddelerine cevap veren katılımcıların sistematik bir biçimde gözlemlenmesini, puanlayıcıların performans görevlerini puanlarken kullandıkları ölçütlerin değerlendirilmesini, madde yanıtlama süresinin analizini (kronometrik analiz), öğrencilerin göz hareketlerinin takibini ve katılımcıların test maddelerini çözerken kullandıkları akıl yürütme sürecinin değerlendirilmesini kapsar. Bu kanıt(lar) özellikle testin üst düzey becerileri ortaya çıkarma derecesini ve farklı alt evrenlerden gelen öğrencilerin test maddelerini ne kadar iyi anladıklarını değerlendirmede oldukça kullanışlıdır. Eğitim alanındaki pek çok yeni testin üst düzey bilişsel becerileri vurguladığı düşünülürse, bu testlerin söz konusu yetenekleri yeterince ölçtüğüne dair kanıtlar gerekecektir.

4. Mutlak değerlendirme ile bağıl değerlendirme yapılan durumlarda geçerlik kanıtlarının toplanmasında farklı yollar izlenmeli midir?

Mutlak ve bağıl değerlendirmenin ne anlama geldiğini anlamadım. Ancak, bir "geçerlik argümanı" ortaya koymak için geçerlik kanıtlarının farklı kaynaklarının nasıl bir araya getirileceği hakkında konuşalım. Kane (1992, 2006, 2013), bir testin belirli bir amaç için kullanımının geçerlenmesinin, testin belirli bir amaç için kullanılmasının doğrulandığı sağlam ve mantıklı bir argüman geliştirmekten farksız olduğunu öne sürmektedir. Standartlar, geçerlik kanıtlarının beş kaynağının testin belli bir amaç için kullanımını desteklemek için tutarlı bir şekilde sentezlenmesi gerektiğini öne sürerek özellikle bu bakış açısını benimsemiştir. Örneğin, "Sağlam bir geçerlik argümanı, mevcut kanıtların ve kuramın belirli kullanımlar için test puanlarının amaçlanan yorumunu destekleme derecesini tutarlı bir açıklamayla birleştirir" diye belirmektedirler (AERA ve diğerleri, 2014, p.21). Geçerlemeye yönelik argüman-temelli yaklaşım, bir mahkeme salonunda bir testin kullanılmasını savunmaya benzerdir. Buradaki fikir, testin belli bir amaç için kullanılmasını destekleyecek çeşitli ve güçlü kanıtlar sunmaktır. Bu kanıtlar bütünü, testin planlanan hedeflerini karşılayan ve istenmeyen sonuçlar doğurmayan kanıtları içermelidir.

5. Bir çalışmada, geçerlik çalışması yapılabilecek yeterli örneklem yoksa geçerlik kanıtı toplamak için ne yapılabilir?

Testin kapsamına dayalı geçerlik kanıtları genellikle alan uzmanları kullanılarak elde edilir ve genellikle çok az sayıda uzmanın ürünüdür (örneğin, 10 veya daha az). Dolayısıyla, geçerlik çalışmalarının tamamen madde yanıtları ve test puanlarının analizine dayalı olmadığını lütfen hatırlayın. Kapsam geçerliği kanıtı eğitimle ilgili testlerde temeldir ve gereklidir. Yanıtlama süreçlerine dayalı geçerlik kanıtına dayanan çalışmaların çoğunluğu da çok küçük örneklem büyüklüğü kullanır.

6. Bir testin cevaplanma süreci geçerlik kanıtı olarak nasıl kullanılmaktadır? Geniş katılımlı sınavlarda cevaplama süreci geçerlik için kanıt olarak kullanılabilir mi?

Üçüncü sorunun cevabına bakınız.

7. Geçerlilik ile çoklu yöntem-çoklu özellik analizi arasındaki ilişki nasıl açıklanır?

Campbell and Fiske'yi (1959) okumanızı öneririm.

8. Boud (1995) ve Messick (1995), alternatif değerlendirme yöntemleri için sonucsal geçerlik kavramını gündeme getirmişlerdir. Değerlendirmenin öğrenme üzerindeki etkisine dayanan sonucsal geçerliğin tayininde hangi yollara başvurulmaktadır?

Messick "sonuçsal geçerlik (consequential validity)" terimini kullanmamıştır. Standartlar, "testin sonuçlarına dayalı geçerlik kanıtlarını" tanımlar. Testin sonuçlarına dayalı geçerlik kanıtı, bir test programı ile ilişkilendirilen kasıtlı (intended) veya kasıtsız (unintended) sonuçların değerlendirilmesini ifade eder. Testler, psikolojik rahatsızlıkların doğru tanılanması, kamunun korunması, öğretimin geliştirilmesi ve yapıların daha iyi anlaşılması gibi pozitif sonuçları arttırmak için kullanılır. Açıkça düşünülmemiş veya planlanmamış kasıtsız pozitif sonuçlar da ortaya çıkabilir. Ancak test programında kasıtsız negatif sonuçlar da ortaya çıkabilir. Belirli gruplar için eğitim ve iş olanaklarında azalmaya neden olan ters etki, okul bırakma oranında artış, test performansına dayalı olarak çalışan maaşları veya kaynak aktarımına ilişkin kötü kararlar kasıtsız sonuçların örnekleri olabilir.

Sonuca dayalı geçerlik kanıtı hesap verilebilirlik (örneğin okul veya öğretmenleri değerlendirmek için testlerin kullanılması) gibi bazı amaçlar için testlerin geçerlenmesi düşünüldüğünde özellikle önemlidir. ABD'de hesap verilebilirliği test etmek federal eğitim politikasınca gereklidir ve genellikle paydaşlar için beklenen sonuçların ana hatlarını belirten bir eylem kuramıyla birlikte gelir. Örneğin, öğrencilerin test sonuçlarına göre öğretmenleri değerlendirmek, amaçlanan müfredatı öğretmek için öğretmenleri teşvik eder ve bu daha çok odaklanmış öğretimin, öğrencilerin söz konusu müfredatı öğrenmelerini geliştireceği varsayılır. İstenen sonuçların gerçekleştirilme derecesi ve istenmeyen sonuçların (örneğin, düşük öğretmen morali, öğrencilerin öğrenmesini azaltacak şekilde müfredatın daraltılması) ne ölçüde en aza indirileceği, hesap verilebilirlik için eğitim alanındaki testlerinin geçerliğinin incelenmesinde önemlidir. Lise mezuniyeti veya üniversiteye giriş için bir test kullanmak gibi, diğer test amaçlarının da değerlendirilmesi gereken sonuçları vardır.

9. Alan yazına bakıldığında da geçerliğe dair farklı yaklaşımlar ve sınıflamalar bulunmaktadır. Durumu özelleştirmek adına hipotetik bir durum üzerinden soru yer almaktadır. -Üst düzey becerileri (ör; problem çözme, eleştirel düşünme) ölçmek istiyorsunuz. Test puanlarının sonucu kritik (high stakes) öneme sahip kararlarda kullanılacak. Geçerliği sağlamak adına testin geliştirilmesinden puanların anlamı ve kullanımına kadar olan süreçte nasıl geçerlik kanıtları sunarsınız?

Dördüncü sorunun cevabına bakınız.

10. Likert tipi ölçeklerde bireylere ait puanların hesaplanması, maddelere verilen tepkilerin sayısal değerlerinin toplanması ile gerçekleştirilmektedir. Fakat her maddenin ilgili yapıya hizmet etme derecesinin farklılık gösterdiği de açıktır. Bu bağlamda, bu gerçeğe rağmen puanların düz bir şekilde toplanması ne derece doğru ve bu puanlardan elde edilen bulguların/ yapılan çıkarımların geçerliliği ne derece yüksektir?

Geçerliğin testin belirli bir amaç için kullanımına karşılık geldiğini hatırlayınız. Genel olarak tek maddelik testler yoktur ve dolayısıyla çıkarımlar ve kararlar pek çok maddeden hesaplanan puanlara dayanarak yapılır. Bir veya iki maddeyle ilgili bir problem varsa bu durum, toplam test puanlarını ve yorumlamayı nasıl etkilediğine bağlı olarak, önemli olabilir de olmayabilir de. Geçerlik analizine madde düzeyi, toplam puan düzeyi ve alt puan düzeyi gibi farklı düzeylerdeki analizleri katmak iyidir. Eğer geçerlik kanıtlarının beş kaynağını dikkate alacaksanız, tüm düzeyler hesaba katılır. Madde analizleri ve değişen madde fonksiyonu içyapıya dayalı geçerlik kanıtının bir parçasıdır. Test kapsamına dayalı çalışmalarda olduğu gibi, boyutluluk analizleri de genellikle madde düzeyine odaklanır. Ancak diğer değişkenlerle ilişkiler (örneğin, yordama geçerliği, değişen yordama geçerliği, çoklu özellik-çoklu metot) toplam test puanı düzeyine odaklanacaktır. Değerlendirilecek olan belli geçerlik sorusu analizin düzeyini belirleyecektir.

11. ***Türkiye olarak sınav uygulamalarında daha çok test güvenliği noktasına odaklanmaktayız ve geçerlik adına ortaya konulan çoğu uygulamaları kurumlarımızın henüz uygulamaya koymadığını görmekteyiz. Diyelim ki Türkiye'de yapılan sınavlar ile ilgili yetkili bir komisyon sizinle çalışmak istedi. Geçerliği bilimsel ve sosyal değerler temelinde düşünerek hangi uygulamaları ne şekilde değiştirmek isterdiniz ve önerileriniz ne olurdu?***

Güvenirlik önemlidir ancak belirli bir amaç için geçerli olmayan testlerin yine de güvenilir olabileceğini hatırlamak daha önemlidir. Örneğin, bir üniversite giriş testi güvenilir puanlar üretebilir; ancak, bir matematik dersinde öğrencilere not vermek için kullanıldığında geçerli olmayacaktır. Sanırım önceki sorular cevaplarım test geçerleme sürecini nasıl ele alacağımı tarif ediyor.

12. ***Türkiye'de ÖSYM'nin (Ölçme-Seçme ve Yerleştirme Merkezi) uyguladığı çeşitli sınavlarda görme engelli adayların görme derecesine göre farklı yazı puntolarında kitapçıklar verildiği, görme yetisi olmayan adaylara ise görsel sorulardan sorumlu tutulmayarak daha az soru sorulduğu, bir okuyucunun desteği ve belirli bir ek süre ile sınava girdikleri bilinmektedir. Bunun yanında çeşitli özür durumları için rapor alan adaylar için de farklı uygulamalar bulunmaktadır (lavabo ihtiyacı gibi). Yukarıda belirtilen benzer durumlar göz önüne alındığında, bir testin geçerlik kanıtları normal şartlar altında mı düşünülmeli veya farklı özelliklere ait alt gruplar için farklı geçerlik kanıtları da aranmalı mıdır?***

Bu geçerlik sorusunun ne olduğuna bağlıdır. Geçerliğin testin belli bir amacına karşılık geldiğini hatırlarsak, testin standart test uygulamaları ve uyarlanmış test uygulamaları için benzer şekilde geçerli olup olmadığı bir soru olabilir. Bir diğer soru, Standart ve uyarlanmış test uygulamalarından elde edilen puanların karşılaştırılabilir olup olmadığı olabilir. Daha özel bir soru ise uyarlamanın ölçülen yapıyı değiştirip değiştirmediği olabilir. Bu alanlarda oldukça geniş bir alanyazın ve uygulamalı araştırma bulunmaktadır.

13. ***Yordama geçerliğindeki ranj daralması sorununa istatistiksel yaklaşımlarla çözüm bulunabilir mi?***

Test verisi değerlendirilirken ranjın daralması gerçekten önemlidir. Muhtemelen değişen ranj daralmasından kaynaklanan gruplar arası değişmezliğin (hem madde hem de toplam test puanı düzeyinde) olmadığı sonucuna varan birçok çalışma gördüm. Madde çift serileri, güvenirlik katsayıları, faktör yükleri vb. gibi istatistikler değişkenliğe ihtiyaç duyar. Eğer bir test bir grup için çok kolay ya da çok zorsa, bu bir yanlılık kaynağı gibi görünecektir ancak bu gerçekte sadece ranj daralmasının yapay etkisidir.

Bu problemle başa çıkmanın bir yolu, sınırlandırılmış grubun dağılımı ile örtüşecek şekilde, sınırlandırılmamış gruptan örneklem çekmektir. Bu strateji değişen ranj daralmasını kontrol altına alır. Tabii ki, ranj daralmasının hafifletilmesi bunu yapabilecek veriye sahipseniz ve uygun düşüyorsa kullanışlıdır.

14. ***Messick (1995) geçerliğin, hem kanıt toplama hem de puan yorumlarının sonuçları ve kullanımı olarak geniş bir biçimde tanımlanabileceğini belirterek kapsam ve ölçüt dayanaklı geçerliği, yapı geçerliğinin alt bölümleri olarak görmüştür. Messick'in birleştirilmiş (unitary) geçerlik kavramı, testlerin geçerliği için toplanan tüm kanıtların testin yapı geçerliğini ortaya koyacağını belirtmektedir. Birleştirilmiş geçerlik kavramı, testin neyi ölçtüğüne dair basit soruları yanıtlamadığı, daha çok nomolojik ağ ile açıklanan karmaşık puan yorumlamalarıyla ilgili olduğu yönünde eleştiriler almıştır. Farklı geçerlik türlerinin farklı amaçlara hizmet ettiği düşünüldüğünde geçerliğe ilişkin (yapı geçerliğinden farklı olarak) bir çatı kavram oluşturmak mümkün müdür?***
15. ***Messick (1995) geçerliğin, hem kanıt toplama hem de puan yorumlarının sonuçları ve kullanımı olarak geniş bir biçimde tanımlanabileceğini belirterek kapsam ve ölçüt dayanaklı***

geçerliği, yapı geçerliğinin alt bölümleri olarak görmüştür. Buna göre kapsam ve ölçüt dayanaklı geçerlikten bağımsız olarak yapı geçerliğini araştırmak ne derece doğru olur? bir ölçümün geçerliği her zaman bir bütün olarak mı araştırılmalıdır? Öncelikle hangi geçerlik türüne bakmak gerekir?

14 ve 15. Lütfen bu iki soru için Sireci (2012)'ye bakınız.

KAYNAKÇA

APA (1952). *Committee on Test Standards. Technical recommendations for psychological tests and diagnostic techniques:* preliminary proposal. American Psychological Association Washington DC US. [American Psychologist, 7(8), 461-475. http://dx.doi.org/10.1037/h0056631]

APA (1954). *Technical recommendations for psychological tests and diagnostic techniques.* American Psychological Association Washington DC US. [Psychological bulletin, 51, 2, pt. 2, March 1954. Supplement.]

APA (1966). *Standards for educational and psychological tests and manuals.* American Psychological Association Washington DC US. [Educational and Psychological Measurement, 26, 3, 751-767 October 1, 1966 DOI: https://doi.org/10.1177/001316446602600328]

APA, AERA, & NCME (1974). *Standardsfor eductiond andpsychological tests and manuals*. Washington, DC: American Psychological Association.

AERA, APA & NCME (1985). *Standards for education and psychological testing*. Washington, DC: American Psychological Association.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Campbell D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait multimethod matrix. *Psychological Bulletin, 56*(2), 81−105.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. http://dx.doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17– 64). Westport: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749

Sireci, S. G. (2012). De-"Constructing" Test Validation. *Center for Educational Assessment Research Report No. 814*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.