Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *İlkbahar* 2018 *Cilt: 9- Sayı: 1*
Spring 2018 *Volume: 9- Issue: 1*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

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

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

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

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)
Akif AVCU (Marmara Üni.)
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.)
Ceylan GÜNDEĞER (Hacettepe Üni.)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Hacettepe Üni.)
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 (Niğde Ömer Halisdemir 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.)
Gül GÜLER (İstanbul Aydın 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.)
İlhan KOYUNCU (Adıyaman 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 Üni.)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
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)
Tuğba KARADAVUT AVCI (Kilis 7 Aralık Üni.)
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 2018, Sayı: 9-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2018, Issue: 9-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);1-16

Development and Examination of Personal and Social Responsibility Behaviors Scale*

Bireysel ve Sosyal Sorumluluk Davranışları Ölçeği'nin Geliştirilmesi ve İncelenmesi

Bijen FİLİZ ** Gıyasettin DEMİRHAN ***

Abstract
In this study, "Personal and Social Responsibility Behaviors Scale (PSRB-S)" was developed in order to determine students' responsibility behaviors in accordance with "Personal and Social Responsibility" model developed by Don Hellison and students' personal and social responsibility levels were examined in terms of gender, age and years of sport practice through this scale. Pertaining to personal and social dimension of responsibility, four-category Likert type trial scale consisting of 52 items and Exploratory Factor Analysis (EFA) were applied to 330 high-school students. Items that did not apply as a result of the analysis were omitted from 52-item trial scale and the scale was reduced to 14 items. A final scale consisting of two factors was created. Obtained scale was applied to different 250 high-school students for Confirmatory Factor Analysis (CFA). It has been determined that EFA and CFA results of two-factor PSRB-S and reliability and validity of internal consistency coefficients are at an acceptable level. It was not detected a significance difference in total scores of athlete students' responsibility behaviors in terms of gender and age variables while there were significant difference in their total scores of years of sport practice.

Keywords: Personal responsibility, social responsibility, physical education and sport

Öz
Bu çalışmada, Don Hellison tarafından geliştirilen "Bireysel ve Sosyal Sorumluluk" modeline uygun olarak öğrencilerin sorumluluk davranışlarını belirlemek amacıyla "Bireysel ve Sosyal Sorumluluk Davranışları Ölçeği (BSSD-Ö)" geliştirilmiş ve bu ölçek aracılığıyla cinsiyet, yaş ve spor yapma yılı değişkenlerine göre bireysel ve sosyal sorumluluk düzeyleri incelenmiştir. Sorumluluğun bireysel ve sosyal boyutuna ilişkin 52 maddeden oluşan dört kategorili Likert tipli denemelik ölçek, Açımlayıcı Faktör Analizi (AFA) için 330 lise öğrencisine uygulanmıştır. Analiz sonucunda 52 maddeden oluşan denemelik ölçekten işlemeyen maddeler çıkartılarak 14 maddeye indirilmiş ve iki faktörden oluşan nihai ölçek oluşturulmuştur. Elde edilen ölçek doğrulayıcı faktör analizi (DFA) için farklı 250 lise öğrencisine uygulanmıştır. İki faktörlü olan BSSD-Ö'nün AFA ve DFA sonuçları ve iç tutarlılık katsayılarının güvenirlik ve geçerliliklerinin kabul edilebilir düzeyde olduğu belirlenmiştir. Sporcu öğrencilerin sorumluluk davranışlarında; cinsiyet ve yaş değişkenlerine göre toplam puanlarında anlamlı bir farklılık bulunmamış, spor yapma yılı toplam puanlarında ise anlamlı farklılıklar bulunmuştur.

Anahtar Kelimeler: Bireysel sorumluluk, sosyal sorumluluk, beden eğitimi ve spor.

INTRODUCTION

Responsibility is an important behavior that every individual should have. The individual begins to develop personal and social responsibility behaviors, which they began to gain in his / her family, in a planned way in the school environment. With the correct responsibility education given in the schools,

* The scale development section of this study was presented as oral presentation on November 12th, 2015 in Morocco/Rabat.
** Dr, Hacettepe University, Faculty of Sport Science, Ankara-TURKEY, bijenfiliz@gmail.com, ORCID ID: https://orcid.org/0000-0001-5863-3861
***Prof. Dr, Hacettepe University, Faculty of Sport Science, Ankara-TURKEY, demirhan@hacettepe.edu.tr, ORCID ID: https://orcid.org/0000-0002-0758-3427

DOI: 10.21031/epod.332978
Geliş Tarihi: 06.08.2017
Kabul Tarihi: 14.12.2017

it is aimed for individuals to become citizens who are aware of their responsibilities and has the ability to increase the protective factors and durability of the individual's life skills.

Responsibility can be defined as "The individual undertaking his/her own behavior and responsibilities of the event falling within the jurisdiction both morally and legally" (Jenkins, 1994), "Starting from early childhood, performing tasks in accordance with the child's age, gender and level of development" (Yavuzer, 2006), "Making choices and accepting the consequences and influences of these choices" (Yalom, 2001). Personal responsibility is "to fully accept all responsibilities and assignments in order to identify and achieve clear goals in life" (Nelson et al., 2004). Social responsibility can be defined as "The person's care for others, fulfilling their obligations to others, participating in the social process, dedication to relieving pain, and endeavoring for a better world" (Lickona, 2009). In this context, creating awareness of both social and personal responsibilities while improving life skills will contribute positively to the person becoming a full individual.

Teaching life skills addresses the emotional and social aspects of being a complete individual. Teaching children life skills because of these and many other reasons is significant, despite the challenges and means helping students to take personal and social responsibility, sharing authority with the students and giving the power to decide them over time (Hellison, 2014, p.13). In this context, according to the Teaching Personal and Social Responsibility (TPSR) literature, it seems that TPSR is the most pressing and typical model in terms of the observed improvement of physical education and sport in the social responsibility behavior (Hellison & Walsh, 2002).

The TPSR model of Don Hellison (2011) which is a program attracting great interest was originally designed for young people at-risk. This model is also known as the responsibility model. In the USA, it has been used in a wide range (Hellison & Walsh, 2002). It is a common education program used in physical education classes, summer and in after-school programs especially for children in at-risk groups that do not get adequate services (Hellison, 2011; Hellison & Martinek, 2006; Hellison & Walsh, 2002).

TPSR training aims to instill the character traits such as social responsibility, taking responsibility for children through physical activities that emphasize the value guidance, and to provide a holistic self-development in gaining basic values (Hellison, 2014).

TPSR is a responsibility-based program which can be used as both a preventative measure and an intervention to support a value and belief system that supports prosocial behaviors in children. It has the ability to increase the protective factors and resiliency of participants who are at-risk for negative outcomes due to their environmental circumstances (Martinek & Hellison, 1997). It aims to empower children to take control of their lives by providing them the chance and space to exhibit responsible behaviors. The program gradually shifts responsibility from the facilitator to the program participants (Hellison & Martinek, 2006; Hellison, 2011; Hellison & Walsh, 2002). Hellison and Martinek (2006) indicates the responsibility of this model's overarching objective as to help the children's development in order to contribute to their own welfare and the welfare of others. The most appropriate environments for gaining responsibility behavior of children in upbringing period are schools. But the schools fail to teach how to achieve a successful identity in terms of social responsibility and self-confidence needs. However, social responsibility training should be a part of each school's program. Otherwise, many children cannot develop a successful personality (Glasser, 1999). The use of technical approach, strategy or model developed according to the field in bringing the children responsible behavior in school may provide, in this regard, a more effective learning environment in schools.

When considering the physical education and sports, it can be said that not all but some scientists respond to these developments. Examples include physical education and physical activity in adventure training (Hattie et al., 1997; Hellison, 2011), character development (Beedy & Zierk, 2000), cooperation (Bressan, 1987; Orlick, 1978), moral development (Gibbons, Ebbeck, & Weiss, 1995; Romance, Weiss, & Bokoven, 1986; Shields & Bredemeier, 1995), good sporting behavior and fair play (Gibbons, Ebbeck, & Weiss, 1995; Giebink & McKenzie, 1985; Horrocks, 1977), empowerment

(Ennis et al., 1999; Siedentop, 1994) and social responsibility (Horrocks, 1978; Trulson, 1986) as well as sports-based youth development programs.

Conceptualization and implementation of these kinds of programs are difficult because personal and social development involves "soft skills", value orientations and intentions, and attitudes as well as specific behaviors. How someone feels- an intangible mix of perceptions and intentions toward the self or someone else- may have greater personal and social implications than more visible behaviors (Hellison, 2014, p.12). Therefore, while teaching life skills, getting children to comprehend and practice the behavior that can/cannot be observed is important for developing sense of responsibility in children.

When looking at the research and the literature related to the liability issues, it is observed that responsibility is divided as the personal and social responsibility (Nelson et al., 2004), and is defined as feelings (Berkowitz, 1963), skills (Chamberlin, 1994; Ellenburg, 2001), and personal characteristics and character (Yalom, 2001). Furthermore, it is stated that there is a close relationship between some variables and responsibility. The locus of social class and responsibility (Berkowitz & Lutterman, 1968; Chebat, 1968) is positively correlated to empathy and happiness (Barrio, Aluja, & Garcia, 2004), and academic success (Taylı, 2006; Golzar, 2006), and it was found internal locus of control and academic success of the individuals were high (Önal, 2005; Taylı, 2006; Golzar, 2006). With responsibility education program, it has been determined that responsibility is a teachable behavior (Glasser, 2005; Önal 2005; Taylı, 2006).

Considering the scale developed related to responsibility in Turkey and abroad; it appears that Responsibility Scale was developed by Golzar (2006) for the fifth grades. Personal and Social Responsibility Scale developed by Conrad and Hedin (1982) was adapted into Turkish by Taylı (2006). Social Responsibility Scale development work of Onal (2005) on high school students is available. Internal and External Supervisory Responsibility Scale (Ozen et al., 2002), and sense of Responsibility and Behavior Scale were developed by Ozen (2013). In studies carried out in stages by Coles and Schofield (2008), they have developed The Pathways to Inflated Responsibility Beliefs Scale to identify ways to increase the personal's responsibility for his beliefs based on the responsibility description of Salkovskis et al. (1999). İkiz et al. (2013) have adopted the scale into Turkish. Patrick et al. (1997) developed the 11-point Social Goal Scale to measure social relations and social responsibility targets of fifth grade students in the classroom. Anderman and Anderman (1999) developed another on a 17-item Social Goal Scale on the scale of Patrick et al. (1997).

When the scales developed on the responsibility for physical education and sports in the international field were analyzed, it is observed that Watson et al. (2003) developed the Contextual Self-Responsibility Questionnaire to be used in physical activity after examining the TPSR model of Don Hellison (2011); and Li et al. (2008) developed personal responsibility and social responsibility component and two-factor 14-item Personal And Social Responsibility Questionnaire by modifying the scale of Watson et al. (2003). Filiz and Demirhan (2015) adapted the Turkish version of the same scale. In addition it is observed that Guan, McBride, and Xiang (2006) carry out studies in physical education and sports programs with high school students in order to examine whether Social Goal Scale developed by Patrick et al. (1997) could be generalized to physical education settings. At the end of the study adapting from the scale of Patrick et al. (1997) developed the 11-point scale Social Goal Scale-Physical Education (SGS-PE).

The school is the most important institution in gaining the responsibility behavior after family (Akyüz, 1991, s. 247). Teachers try to develop responsibility by giving students small responsibilities for their behavior in the learning area and, latent and formal messages in teaching activities. They use authorization, reflection time and group meetings techniques in the assessment of responsible behavior; for informal student assessment, based on the degree of personal and social responsibility to give feedback they receive from the course; for the official student assessment use logging, the scoring key, give their student opportunity to mention scoring techniques (Hellison, 2014). Diagnostic tools are required in order to grade students' responsibility behaviors.

Purpose and importance of research according to the relevant literature, it has been seen that there is only one scale related to responsibility behavior in the field of physical education and sports in Turkey and that there is a limited number of scales in the world. In addition, it has been considered appropriate to improve this scale as it shortens the time that teachers spend for grading, simplifies the evaluation of student work, as it brings about an observable criterion and in terms of getting a feedback related to the impact of education on forming responsibility behaviors. The purpose of the study within this scope is to develop a scale in conformity with personal and social responsibility model in an attempt to determine students' responsibility behaviors in the field of physical education and sports; and to analyze students' level of responsibility in terms of gender, age and years of sports practice by evaluating the data obtained from the scale.

METHOD

Research Design

In this study, it was aimed to investigate the level of responsibility of the students in terms of gender, age and years of sport practice using the Personal and Social Responsibility Behavior Scale developed by the researcher. In this study, a relational screening model was used to determine the level of personal and social responsibility of the students (aimed to present the current situation).

Study Group

Study group of Exploratory Factor Analysis (EFA) consists of 330 high school students in total, 128 female, 202 male, studying in high schools which are related to Ministry of Education (MEB) and pursue their educational activities in 2015-2016 academic year in Yenimahalle district center of Ankara (TVF Sport High School: n1=120, 60 boys, 60 girls; Gazi Technical and Industrial Vocational High School: n2=50, 50 boys; Gazi Ciftligi Anatolian High School: n3=80, 40 boys, 40 girls; Mimar Sinan Technical and Industrial Vocational High School: n4=80, 52 boys, 28 girls). Demographic information of participant students is presented in Table 1.

Table 1. Demographic Information of Athlete Students for EFA

Variable	Group	n	%	Total
Gender	Female	128	38.8	330
	Male	202	61.2	
Age	14-15	112	33.9	330
	16-17	218	66.1	
The years of sport practice	1-2 years	64	19.4	330
	3-4 years	115	34.8	
	5-6 years	106	32.1	
	7-8 years	45	13.6	

Study group of Confirmatory Factor Analysis (CFA) consists of different 250 high school students in total, 93 female, 157 male, studying in high schools which are related to Ministry of Education (MEB) and pursue their educational activities in 2015-2016 academic year in Yenimahalle district center of Ankara (TVF Sport High School: 57 boys, 53 girls; Gazi Technical and Industrial Vocational High School: 60 boys; Atatürk Anatolian High School: 40 boys, 40 girls). Personal information of participant students is presented in Table 2.

Table 2. Demographic Information of Athlete Students for CFA

Veriable	Group	n	%	Total
Gender	Female	93	37.2	250
	Male	157	62.8	
Age	14-15	109	43.6	250
	16-17	141	56.4	
The years of sport practice	1-2 years	35	14.0	250
	3-4 years	94	37.6	
	5-6 years	75	30.0	
	7-8 years	46	18.4	

Measurement Instrument

Personal and social responsibility behaviors scale (PSRB-S)

PSRB-S, developed by the researcher, was used in the study. Scale items were prepared by utilizing the scale developed items of Li et al. (2008) (1, 2, 3 and 8. items) and from the TPSR model of Hellison (2011) contemplated as the most appropriate model in conferring the behavior responsibility in physical education and sport. Examining textbooks and items related to the model (Hellison, 1976; Hellison, 1978; Hellison, 1985; Hellison, 2011; Hellison & Cutforth, 2000; Hellison & Walsh, 2002), a total of 56 items has been established consisting of three negative (personal responsibility), 53 positive items in personal areas 28 and 28 agents in the social sphere for the scale. The scale was initially applied by the researcher to a student group of 30 in order to test the understandability of items for students. One of the commonly used methods of determining the validity that expresses for measuring the quality and quantity of the desired properties of the items used is to apply to expert opinion (Büyüköztürk, 2016). Items obtained within this frame were reviewed by five academician experts in fields of psychological consultancy and guidance, educational program, education, physical education and sports psychology and four items were removed out of the scale as a result of experts' opinions and students' feedbacks and a revised 52-item form was created. Afterwards, the scale was applied by the researcher to 340 students in class environment by visiting random classes in four different high schools. Students were given 25 minutes by the researchers in order to fill 52-item scale form. At the end of the application, test forms which are not suitable for the validity and reliability studies were removed and EFA was conducted on 330 students in total, 128 female and 202 male. Items that did not load on a certain factor as a result of the analysis were omitted from 52-item trial scale and the scale was reduced to 14 items and a final scale consisting of two factors was created.In this study, it has been avoided to make the mistake of using third degree 'Neither Disagree, nor Agree' and 'No Opinion' expressions which are presented in 5-degree scales (Bohner and Wanke, 2002, Şencan, 2005). It has been decided to form the distracters in four categories from Never (0) to Always (3) and to make scoring between 0 and 3 with regard to the responses given to the scale items to be less distractive in terms of evaluating responsibility behavior. In this way, it was aimed to receive reliable answers related to the situations where students avoid to express their opinion, in other words choosing the option 'Neither Disagree, nor Agree', by taking a step towards determining their actual tendencies. The highest possible score in scale is 56 and lowest is 14. According to this grading, 0-14 point in personal and social responsibility scale can be regarded as negative, 15-28 points as average, and 29-42 points as positive and 43-56 points as highly positive.

The structure of the social responsibility of the scale represents two of the TPSR levels: Respect for the rights and feelings of others and helping/ leadership others. As a sample item "I show sensitivity to the skill level of my friends in the group work." can be showed. The structure of personal responsibility represents the two levels of TPSR: Effort/ participation and self-direction. As a sample item "I prepare my work plan according to my personal needs." can be given. When the dispersion of the material is examined; it seems that to show respect for the rights and feelings of others' is 4 items, effort/ participation is 3 items, self-direction is 4 items, helping/ leadership others is 3 items. There were questions regarding gender, age and years of sport practice in personal information section of the scale in order to collect the data to be used as independent variable.

Data Analysis

SPSS 20.0 package program and LISREL 8.80 were used in data analysis. For the validity of the scale, firstly homogeneity of the scale scores was checked, then Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were conducted (Büyüköztürk, 2016). Correlations of the items, which are in sub-dimensions of the scale, were analyzed in order to provide evidence for validity of the items. Cronbach alpha coefficient was calculated for internal consistency of the scale. Pearson Correlation analysis was performed in an attempt to determine the linear relation between two sub-factors. Firstly, homogeneity of scale scores was checked in order to compare students' scores from the scale according to independent variables; then t test was performed between two independent groups; one way variance of analysis was used in order to test the difference between the average of groups more than two (Büyüköztürk, 2016; Field, 2005). As significance level .05 was taken in analysis and interpretation of the data.

RESULTS

Results of the research were presented in two parts as the findings aimed at validity and reliability studies that were conducted concerning development of the scale and the findings aimed at examination of students' responsibility behaviors in terms of certain variables.

Results Pertaining to Development of the Scales

Validity study of the scale

Firstly, normality distribution of the scale was reviewed and it was determined that students' scores out of the scale is between -1.5 and +1.5 skewness and kurtosis ranges; it was observed that the data have normal distribution (Tabachnick & Fidell, 2013).

Exploratory factor analysis (EFA): In order to determine the construct validity of the scale was performed the EFA on the data. Before the EFA, three negative items were reversed and then reliability analysis was performed. Five items that corrected item total correlation values is below .30 as a result of the analysis were removed from the scale (30, 31, 33, 41, and 51). Bartlett test was found to be significant as a result of analysis's principle component before rotation. Sample size conducted in order to determine eligibility to factoring KMO value was determined to be .90. KMO value according to the relevant literature middle .60, .70 is good, very good .80, .90 is considered excellent (Bryman & Cramer, 1999). Therefore, the approach to one of the KMO value (.90), the sample size of it and Bartlett test was excellent to reveal the existence of the correlation between the scale items results indicate that suitable for factor analysis of the data sets obtained. Applied Bartlett test results obtained Chi-square test statistics were significant ($\chi 2$= 4925.0456; df= 1081, p<.01).

Before rotating the factors, 12 factors with eigenvalues higher than 1.00 were revealed. These factors explain the % 56.664 of the variance related to responsibilities variable. Varimax rotation technique (Varimax component analysis) was used to group the personal and social items for EFA. It is mentioned that on scale development regarding the creation of factors which could then be taken as the lower cut-off point of factor loadings are ranging from .30 to .45 (Büyüköztürk, 2016). When the distribution of the factors load was examined, it was discovered that scale items were tend to gather under two factors. By taking the breakpoint as .40, 33 items (1, 2, 4, 5, 6 , 9, 10, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 32, 35, 36, 37, 39, 40, 43, 44, 45, 46, 47 , 49, 50, 52) that the factor loadings are below .40 and showing scattered on various factors (complex) were removed from the scale. After that there are 14 significant items. When the analysis was repeated with the removed item factors, only 2 factors with eigenvalues higher than 1.00 are found. As seen in Table 3, when we look at the distribution of the factor load; seven items formed by the first sub-factors (27, 28, 29, 34, 38, 42, and 48) and the eigenvalue of the first factor that information on the significance and weight of each factor in the structure was found to be 4.254. The responsibility variable factor alone explains the 30,387 %.

The second component of seven substances (3, 7, 8, 11, 12, 13, and 26) formed and was found to be 1.507 eigenvalue of this factor. The variance of this component alone explains the responsibilities variable 10.761 %. These two factors explain the lower 41,148 % of the variance with the related responsibilities variables. This result, Kline (2011) as indicated by the acceptable limit is above 41%.

In order to provide evidence for the validity of the substance of the scale, the correlations between each item in the sub-dimensions have been analyzed. Accordingly, it has been seen that all of the items are correlated with the .01 level of significance in the medium and high-level (p<.01). The factor loads of the scale have been calculated as principal component analysis, Varimax component analysis and corrected item-total correlation. In Table 3, the results of this analysis are given:

Table 3. Factor Loadings for Exploratory Factor Analysis

Factor name	Item	New Item	Items (Levels of responsibility)	Factor load values	Varimax component factor load values		Corrected item total correlation
					PR	SR	
	3	1	I try the given new tasks (Effort/ participant)	.48	.65		.40
	7	2	I participate in all of the activities (Effort/ participant)	.44	.58		.36
	8	3	I give effort to overcome difficult tasks (Effort/participant)	.61	.46		.51
Personal responsibility (PR)	11	4	I perform a given task without peer pressure (Self-direction)	.48	.50		.39
	12	5	I prepare my work plan according to my personal needs (Self-direction)	.48	.68		.41
	13	6	I do independent study related to my skill level without directed by someone else (Self-direction)	.40	.53		.30
	26	7	I follow the necessary rules to fulfill my responsibilities (Self-direction)	.59	.50		.49
	27	8	I respect others (Respect)	.65		.74	.52
	28	9	I respect my teachers (Respect)	.58		.75	.45
Social	29	10	I control my behavior towards others (Respect)	.62		.75	.49
responsibility	34	11	I care about others (Respect)	.58		.61	.47
(SR)	38	12	I show sensitivity to the skill level of my friends in the group work (Helping/ leadership)	.62		.49	.51
	42	13	I would help others while learning something new (Helping/ leadership)	.59		.48	.48
	48	14	I would help immediately when others ask for help (Helping/ leadership)	.58		.56	.47
			Eigenvalues		1.507	4.254	
			Explained the total variance		10.761	30.387	

Examining Table 3; it is seen that all principal factor load values are higher than .40. On the other hand, it has been determined that Varimax component factor load values are high for all factors and the lowest one is .46. Examining the values on "total item correlation" column giving the correlation of items that constitute the scale with the entire scale; it is seen that the lowest correlation is in the 13. item and in level .30. Thus, it is required for these values to be above .20, which has been provided.

When the items gathered under two factors as a result of the analyses have been examined, it has been determined that the items under the first factor evaluate the social dimension of responsibility and the items under the second factor evaluate the personal dimension of responsibility. According to these items in the first factor is named social responsibility, items in the second factor personal responsibility. Two-dimensional structure of personal and social responsibility scale was found to be related with the conceptual framework related to responsible behavior (Nelson et al. 2004; Golzar, 2006; Hellison, 2011). For this reason, it is considered that the structure of the scale should be retained conceptually.

Confirmatory factor analysis (CFA): In order to reach goodness of fit values of the two-factor model, CFA was performed using the LİSREL 8.80 program with a data group of 250 different high school students (Jöreskog & Sörbom, 2004). Figure 1 shows the diagram of the model.

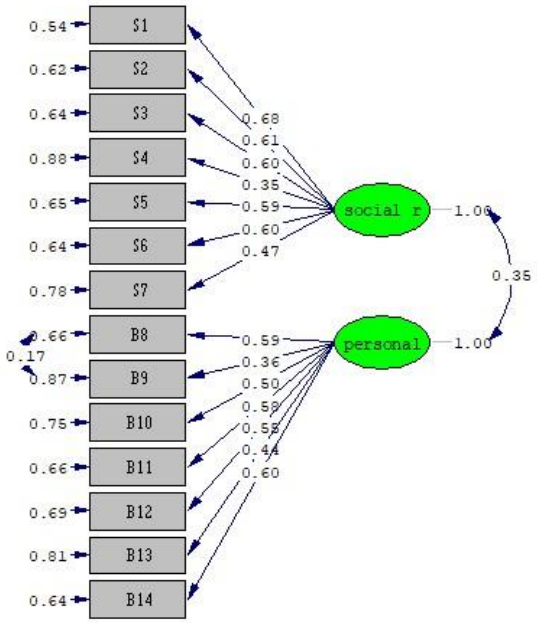


Figure1. Path Diagram of the Model

Examining Figure 1; it is observed that path coefficients between items and their dimensions vary between .35-.68 for social responsibility and .36-.60 for personal responsibility. These items explain at least .54 variance. Because the variance they explain and relation values are moderate and above, these values are accepted as sufficient (Büyüköztürk, 2016). As is seen in Figure 1; error variances of eighth and ninth items in the dimension of personal responsibility are equalized since they reduce the chi-square value. Examining the contents of these items; the eighth item is as, "I try new tasks", whereas the ninth item, "I participate in all activities". Both items are statements that complete each other and seem to be parallel regarding personal responsibility. Thus, it may be concluded that equalization of error variances in CFA is a convenient procedure."

Table 4. Goodness of Fit Index Values of CFA

X^2	*Sd*	x^2/sd	*P*	GFI	CFI	NFI	NNFI	SRMR	RMSEA	%90 C.J. RMSEA
78.87	75	1.05	0.0	.96	.99	.93	.99	.027	.014	.0-.040

Examining the measurement values of the confirmatory factor analysis as in Table 4; the Chi-square value concerning the 4-item two-factor model was determined as $\chi 2$ (75, n=250)=78.87, p<.001. As a result of the calculation, it was observed to have a good value as $\chi 2/sd$=1.05. It was determined that fit values of RMSEA=.014, SRMR=.027, CFI=.99, NFI=.93, NNFI=.99, GFI=.96, which are frequently used in CFA measurements, had good and excellent fit values (Hu & Bentler, 1999). It was also determined that the confidence interval (CI) limit of 90% varied between .0-.040 for RMSEA. As RMSEA and SRMR values are smaller than .80, the model is considered acceptable (Anderson &

Gerbing, 1984; Cole, 1987). Findings acquired from the confirmatory factor analysis signify that factor structure of the scale shows an acceptable compatibility with the collected data.

Reliability study of the scale

Internal consistency coefficients were calculated to determine the reliability of the scale. Cronbach's Alpha coefficient for these two factors of the scale was .82; for the first sub-factor Cronbach's Alpha coefficient was .78, for the second sub-factor Cronbach's Alpha coefficient was found as .70. When the correlation between the two factors was examined, it was found that there was a significant correlation between the factors. The correlation coefficient of .90 is stated as near perfect, .80 very well, .70 near enough, that the higher than .60 it is dependent on the availability of sizes and it is stated they all together measure a single conceptual structure, below .50 as insufficient (Şencan, 2005; Kline, 2011). High and statistically significant correlation coefficients indicate that the two sub-factors are responsibility components.

The Pearson Correlation Analysis was conducted for testing the relationship between items separated into two sub-factors. As a result of the Pearson Correlation Analysis that was conducted by averaging the items in the two sub-factors with 95% confidence, it was determined that there was a positively significant linear relationship between the items of personal responsibility and social responsibility. [$r(330)= .516$; $p<.01$]. Accordingly, it may be interpreted that in the phase of developing both personal and social aspects of responsibility while gaining responsibility behaviors, in case that there is a parallel positive increase in both dimensions or failure of development, both dimensions may be affected negatively. The student's sense of personal responsibility is high so social responsibility behavior is also high or can be expressed vice versa.

Results Pertaining to Students' Responsibility Behaviors

In this section is contained the findings concerning whether students display responsibility behaviors according to variables of gender, age and years of sport practice. Firstly, normality distribution of the scale was reviewed and it was determined that students' scores out of the scale is between -1.5 and +1.5 skewness and kurtosis ranges; it was observed that the data have normal distribution (Tabachnick & Fidell, 2013).

Table 5. Differences Between Mean Scores of Responsibility Behaviors by Gender Variable (t test)

Dimension	Gender	N	$\overline{X}$	Sd	t	p
Personal responsibility	Female	128	20.86	3.25	.748	.45
	Male	202	20.58	3.26		
Social responsibility	Female	128	23.18	3.66	1.787	.07
	Male	202	22.46	3.46		

$p<.05$

When Table 5 was reviewed, it was not found a significance difference between male and female's mean scores based on dimensions of responsibility of gender ($p>.05$).

Table 6. Between Mean Scores of Responsibility Behaviors by Age Variable (t test)

Dimension	Age	N	$\overline{X}$	Sd	t	P
Personal responsibility	14-15	112	20.82	3.05	.522	.60
	16-17	218	20.62	3.35		
Social responsibility	14-15	112	23.02	3.59	1.010	.31
	16-17	218	22.60	3.53		

$p<.05$

When Table 6 was reviewed, it was not found a significance difference between 14-15 and 16-17 ages mean scores based on dimension of responsibility of age (p>.05).

Table 7. Differences Between Mean Scores of Responsibility Behaviors by The Years of Sport Practice Variable (One Way ANOVA)

Responsibility behavior	Source of variance	Sum of squares	Df	Mean square	F	P	Difference
Personal responsibility	Between groups	23.923	3	7.974			
	Within groups	3458.55	326	10.609	.752	.522	
	Total	3482.47	329				
Social responsibility	Between groups	97.654	3	32.551			7-8 *5-6 years
	Within groups	4009.24	326	12.298	2.647	.049*	
	Total	4106.89	329				

*P<.05

Examining Table 7; it was determined that there was a significant difference between the score averages of answers given by students to judgments regarding their perception on the sub-dimension of social responsibility [F(3,326)=2.647, p<.05] according to the variable of years of sport practice, whereas there was no difference on the sub-dimension of personal responsibility [F(3,326)=.752, p>.05].

Evaluating the students according to years of sport practice in the results of the multiple comparison test; it is seen that there is a significant difference between 5-6 years and 7-8 years based on the sub-dimension of social responsibility (p<.05). In social responsibility sub-dimension, point average of the answers given to 7-8 years is 21.400, point average of answers given to 5-6 years is 23.141.

DISCUSSION

First, in this study, in the implementation of the TPSR model, in order to evaluate the responsibility behavior of students and athletes studying at high school PSRB-S was developed. Scale items are consistent with the conceptual model of TPSR literature (Hellison & Martinek, 2006; Hellison & Walsh, 2002). As a result of the analysis, it was determined that there is a positive relationship between personal and social responsibility structures of the scale. Findings regarding validity and reliability that were acquired as a result of the study indicate that the scale is convenient for determining behaviors aimed at the attributes in question.

PSRB-S can be used in different age groups like primary, secondary and high school students, physical education and sport activities at schools, club activities, after school programs and camp activities for evaluating responsibility behaviors and this scale can contribute to studies related the field. As a result of the implementation of the scale on students and athletes in different age groups it is expected to reach similar findings in relation to the validity and reliability. The scale was applied to high school students and the validity and reliability analysis of data obtained from this group have been made. If the scale is used to determine responsibility for the behavior of different age groups, it is recommended that it is used after with validity and reliability of data derived from the group performed again.

Although the scale was developed to be used in applications of TPSR model, it can contribute in areas such as guidance, psychology and training involving activities related to responsible behavior. It is considered that the scale would be useful to strengthen the work of the researchers and teachers.

The scale can be more effective when supported by other tools besides course applications. Various tools like self-assessment, graduated scoring, portfolio, rubric formation and taking daily notes are used in model applications. It should not be considered that scale is useful alone without applications that enable gaining behavior and without supported with other diagnostic tools.

It was adhered to the values that contain levels in TPSR model while developing scale items (respect for the rights and feelings of others, effort/ participation, self-direction and helping/ leadership others). In the future, another study can be carried out by considering values such as cooperation, trust, honesty, self-efficacy that have been formed in the process of application in the TPSR model and the number of assessment tools can be replicated relevant to responsibility. In this scale, four-point Likert-type scale was used; in future studies scale can be prepared of five, six or seven-point Likert-type. The scale items can be used in the studies of other areas and in the evaluation of responsible behavior in different sports.

Second subject of analysis in this study was athlete students' personal and social behaviors according to variables of gender, age and years of sport practice via PSRB-S. This study revealed significant differences in social responsibility behaviors of athlete students according to years of sport practice. According to the acquired results, it is possible to state that long years of engagement in sports starting in early ages have a negative effect on social responsibility behaviors of athlete students.In differing of sub-dimensions of PSRB-S depending on gender; it was not found a significant difference between point averages of the answers given by athlete students to their personal and social responsibility behavior ($p>.05$). It is possible to construe this situation as male and female athletes in the age group of study were given responsibilities of equal conditions in sports practices in class activities and clubs of schools and there is not sexual discrimination in terms of bringing in responsibility behaviors. In the study that used Escape from Responsibility Scale developed by Powell, Rosen and Huff (1997), it was found that male students escape from responsibilities significantly more than female students (Powell & Rosen, 1998). In a study conducted by Gunnoe, Hetherington and Reiss (1999), it was found social responsibility level of girls higher than boys. In a study conducted by Taylı (2013), it was found that responsibility level of girls is higher compared to boys. In the study that was conducted by Wright (2011) using the standardized questionnaire for the purpose of examining the responsibility behaviors of students regarding environment and recycling, no difference was found between genders. These results display parallelism with the result of this research.

In differing of sub-dimensions of PSRB-S depending on age; it was not found a significant difference between point averages of the answers given by athlete students to their personal and social responsibility behavior ($p>.05$). Development of the sense of responsibility is possible with the steps taken as of the first years of life. Responsibility is a skill that children initially learn from their parents then from social environment (Gordon, 2010). Teaching of the sense of responsibility that starts in the family continues in school. The scale was applied to students between ages of 14-17 practicing sports. Considering that certain responsibility behaviors have been taught to students in family and school until this age range, it is considered appropriate to compare responsibility behaviors with students of smaller age and different age range. In the study conducted by Wright (2011) on university students, it was not found a difference between students' responsibility behaviors regarding environment and recycling depending on the variable of age. This situation was construed as responsibility behaviors might result in differently in different populations. This conclusion shows parallelism with the result of this study.

In differing of sub-dimensions of PSRB-S depending on years of sport practice in social responsibility sub-dimension; students who practiced sports for 5-6 years presented opinions that are more significant compared to students with 7-8 years of sport practice. It is found that students who have been engaged in sports 5-6 years display more responsibility compared to students who practiced sports for than 7-8 years in terms of behaviors of helping others, caring about others, being respectful towards teachers and others and self-control towards the others. This situation can be interpreted as the more students' years of sport practice increase, the less their social responsibility behaviors become. When there is a decrease in their social responsibility behaviors, it is considered appropriate to examine factors such as students' level of exhaustion, their level of unit and solidarity among the team or class, their internal and external motivation levels, level of responsibility education given throughout their sports life. In the study conducted by Eilam and Trop (2012) on children and their parents in order to examine their attitudes towards environment, teachers provided environmental education to children along with their parents and as a result of the comparison at the end of the study, it was found that parents display more positively-oriented responsibility behaviors compared to

children. This situation was construed as when a person matures, the process positively affects behaviors. This study does not support the result of the research.

When research results were reviewed, it was found that personal and social responsibility behaviors of high-school sportsman students do not display difference basing on gender and age and there is a decrease in displaying social responsibility behaviors in students who started practicing sports at an early age and continued for a long time. According to these results; individuals' personal and social responsibility behaviors can be compared among different populations by categorizing them as primary school, secondary school, high school and university students and even adults. Personals' responsibility behaviors can be analyzed with different variables. Personals can be provided responsibility education in different age groups and change of responsibility behaviors between the age groups can be analyzed. The causes behind why the increase in years of sport practice decreases social responsibility behaviors can be researched.

REFERENCES

Akyüz, H. (1991). *Eğitim sosyolojisinin temel kavram ve alanları üzerine bir araştırma*. İstanbul: Milli Eğitim Bakanlığı.

Anderman, L. H., & Anderman, E. M. (1999). Social predictors of changes in students' achievement goal orientations. *Contemporary Educational Psychology, 25,* 21-37.

Anderson, J. C, & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, 49, 155-173.

Barrio, D. V., Aluja, A., & Garcia, L. F. (2004). Relationship between empathy and the big five personality traits in a sample of Spanish adolescents. *Society for Personality Research*, *32*(7), 677-682.

Beedy, J. P., & Zierk, T. (2000). Lessons from the field: Taking a proactive approach to developing character through sports. *CYD Journal: Community Youth Development, 3*, 6-13.

Berkowitz, L. D. (1963). Responsibility and dependency. *Journal of Abnormal and Social Psychology, 66,* 429-436.

Berkowitz, L., & Lutterman, K. (1968). The traditional socially responsible personality. *Public Opinion Quarterly, 32*, 169-185.

Bohner, G., & Wanke, M. (2002). *Attitudes and Attitude Change*. Hove: (UK) Psychology.

Bressan, E. S. (1987). Physical education and social change in South Africa. In M. Carnes & P. Stueck. (Eds), *Proceedings of the fifth curriculum theory conference in physical education* (pp. 128-138). Athens, GA: University of Georgia.

Bryman, A., & Cramer, D. (1999). *Quantitative data analysis with SPSS release 8 for windows.* London and New York: Taylor & Francis e-Library, Routledge.

Büyüköztürk, Ş. (2016). *Veri Analizi El Kitabı*. (22. bs). Pegem Akademi.

Chamberlin, L. J. (1994). Developing responsibility in today's students. *Clearing House, 67*(4), 204-206.

Chebat, J. (1986). Social responsibility, locus of control and social class. *Journal of Social Psychology, 126*(4), 559-562.

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology*, 55, 1019-1031.

Coles, M. E., & Schofield, C. A. (2008). Assesing the development of inflated responsibility beliefs: The pathways to inflated responsibility beliefs scale. *Behavior Therapy, 35*, 322-335. doi: 10.1016/j.beth.2007.09.003

Conrad, D., & Hedin, J. A. (1982). *Executive summary of the final report of the experiential education evaluation project.* St. Paul, MN: University of Minnesota, Center for Youth Development and Research.

Eilam, E. & Trop, T. (2012). Environmental attitudes and environmental behavior: Which is the horse and which is the cart?. *Sustainability, 4*, 2210-2246. doi:10.3390/su4092210.

Ellenburg, F. C. (2001). Society and school must teach responsible behavior. *Educational Administration, 106*(1), 9-11.

Ennis, C. D., Solmon, M. A., Satina, B., Loftus, S. J., Mensch, J., & McCauley. M. T. (1999). Creating a sense of family in urban school using the "Sport for Peace" curriculum. *Research Quarterly for Exercise and Sport, 70*, 273–285.

Field, A. (2005). *Discovering statistics using SPSS.* London: SAGE.

Filiz, B. ve Demirhan, G. (2015). Bireysel ve sosyal sorumluluk ölçeğinin (BSS-Ö) Türk diline uyarlanma çalışması. *Hacettepe Üniversitesi Spor Bilimleri Dergisi, 26(*2), 51-64.

Gibbons, S. L., Ebbeck, V., & Weiss, M. R. (1995). Fair play for kids: Effects on the moral development of children in physical education. *Research Quarterly for Exercise and Sport, 66*, 247-255.

Giebink, M. P,. & McKenzie, T. L. (1985). Teaching sportsmanship in physical education and recreation: An analysis of interventions and generalization effects. *Journal of Teaching in Physical Education, 4*, 167-177.

Glasser, W. (1999). *Başarısızlığın olmadığı okul* (K. Teksöz, çev. ed.). Ankara: Beyaz.

Glasser, W. (2005). *Responsibility, respect and relationships: Creating emotionally safe classrooms.* Chatsworth, CA: Quality Educational Programs, Inc.

Golzar, F. A. (2006). *İlköğretim 5. sınıf öğrencilerine yönelik sorumluluk ölçeğinin geliştirilmesi ve sorumluluk düzeylerinin cinsiyet, denetim odağı ve akademik başarıya göre incelenmesi*. (Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.

Gordon, B. (2010). An examination of the responsibility model in a New Zealand secondary school physical education program. *Journal of Teaching in Physical Education*, *29*(1), 21-37. doi: 10.1123/jtpe.29.1.21

Gunnoe, M. L., Hetherengton, E. M., & Reiss, D.(1999). Parental religiosity, parenting style and adolescent social responsibility, *Journal of Early Adolescence*, *19*(2), 199-225.

Guan, J., McBride, R. E., & Xiang, P. (2006). Reliability and validity evidence for the social goal scale- physical education (SGS-PE) in high school settings. *Journal of Teaching in Physical Education, 25,* 226–238. doi: 10.1123 / jtpe.25.2.226

Hattie, J., Marsh, H. W., Neill, J. T., & Richards, G. E. (1997). Adventure education and outward bound: Out-of-class experiences that make a lasting difference. *Review of Educational Research, 67,* 43-87.

Hellison, D. (1976). *Personalized-learning in physical education.* Washington, NW: AAPHER.

Hellison, D. (1978). *Beyond balls and bats: Alienated (and other) youth in the gym.* Washington, DC: AAHPER.

Hellison, D. (1985). *Goals and strategies for teaching physical education.* Champaign, IL: Human Kinetics.

Hellison, D. (2011). *Teaching personal and social responsibility through physical activity.* (3nd ed.). Champaign, IL: Human Kinetics.

Hellison, D. (2014). *Fiziksel aktivite yoluyla bireysel ve sosyal sorumluluk öğretimi* (Çev.ed. B. Filiz). Ankara: Nobel Akademi.

Hellison, D., & Cutforth, N. (2000). *Youth development and physical activity.* Champaign, IL: Human Kinetics.

Hellison, D., & Martinek, T. (2006). Social and individual responsibility programs. In D. Kirk, M. Macdonald & M. O'Sullivan (Eds.), *The handbook of physical education* (pp. 610- 626). Thousand Oaks, CA: Sage.

Hellison, D., & Walsh, D. (2002). Responsibility-based youth programs evaluation: Investigating the investigations. *Quest, 54*, 292-307.

Horrocks, R. N. (1977). Sportsmanship. *Journal of Physical Education, Recreation and Dance, 48*, 20-21.

Horrocks, R. N. (1978). Resolving conflict in the gymnasium. *Journal of Physical Education, Recreation and Dance*, *49*(7), 61-61.

Hu, L. T, & Bentler, P. M. (1999). Cutoff criteria for fit ındexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling,* 6(1), 1-55.

Jenkins, D.R. (1994) An eight-step plan for teaching responsibility. *The Clearing House*, 67 (5), 269-270.

İkiz, F. E., Tarık, T. ve Karaca, R. (2013). Sorumluluk inançlarını arttıran faktörleri belirleme ölçeğinin uyarlanması. *New Symposium Journal, 51*(2).105-114.

Kline, R. B. (2011). *Principles and practice of structural equation modeling.* (3nd ed.). New York London: The Guilford.

Kumchy, C. I., & Sayer, L. A. (1980). Locus of control in a delinquent adolescent population. *Psychological Reports*, *3*(2), 1307-1310.

Lee, O., & Martinek, T. (2009). Navigating two cultures: An investigation of cultures of a responsibility-based physical activity program and school. *Research Quarterly for Exercise and Sport*, *80*(2), 230-240. doi: 10.1080/02701367.2009.10599557

Li, W., Wright, P. M., Rukavina, P., & Pickering, M. (2008). Measuring students' perceptions of personal and social responsibility and its relationship to intrinsic motivation in urban physical education. *Journal of Teaching in Physical Education*, *27*, 167-178.

Lickona, T. (2009). *Educating for character: How our schools can teach respect and responsibility*. New York: Bantam.

Martinek, T. J., & Hellison, D. R. (1997). Fostering resiliency in underserved youth through physical activity. *Quest*, *49*, 34-49.

Martinek, T., Schilling, T., & Johnson, D. (2001). Transferring personal and social responsibility of underserved youth to the classroom. *The Urban Review*, 33(1), 29-45.

Nelson, D. B., Low, G. R., Stottlemyer, B. G., & Martinez, S. (2004). *Personal responsibility map (PRM).* Appleton, WI: Oakwood Solutions, LLC.
Orlick, T. (1978). *The cooperation book of games and sports.* New York: Pantheon.
Önal, Ş. (2005). *Bir sorumluluk eğitim programının lise dokuzuncu sınıf öğrencilerinin sorumluluk düzeylerine etkisi*. (Yayımlanmamış yüksek lisans tezi, Uludağ Üniversitesi, Bursa). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.
Özen, Y. (2013). Sorumluluk duygusu ve davranışı ölçeğinin geliştirilmesi geçerliği ve güvenirliği. *Gümüşhane Üniversitesi Sosyal Bilimler Elektronik Dergisi*, *7*, 343-357.
Özen, Y., Gülaçtı, F. ve Çıkılı, Y. (2002). İlköğretim öğrencilerinin sorumluluk duygusu ve davranış düzeyleri ile iç-denetimsel sorumluluk ile dış-denetimsel sorumluluk düzeyleri arasındaki ilişkinin incelenmesi. *Erzincan Eğitim Fakültesi Dergisi, 4*(2), 45-58.
Patrick, H., Hicks, L., & Ryan, A. M. (1997). Relations of perceived social efficacy and social goal pursuit to self-efficacy for academic work. *Journal of Early Adolescence*, *17*, 109-128.
Phares, E. J. (1976). *Locus of control a personality determinant of behavior*. Morritown, NJ: General Learning.
Romance, T. J., Weiss, M. R., & Bokoven. J. (1986). A program to promote moral development through elementary physical education. *Journal of Teaching in Physical Education*, *5*, 126-136.
Powell, K. M., & Rosen, L. A.(1999). Avoidance of responsibility in conduct disordered adolescents. *Personality and Individual Differences*, *27*(2),327-340.
Powell, K. M., Rosen, L. A., & Huff, M. E. (1997). Disruptive behavior disorder and the avoidance of responsiblity. *Journal of Personality and Individual Differences*, *23*(4), 549-557.
Salkovskis P, Shafran R, Rachman S., & Freeston M. H. (1999). Multiple pathways to inflated responsibility beliefs in obsessional problems: Possible origins and implications for therapy and research. *Behav Res Ther*, *37*(11), 1055-1072.
Shields, D. L. L., & Bredemeier, B. J. L. (1995). *Character development and physical activity.* Champaign, IL: Human Kinetics.
Siedentop, D. (1994). *Sport education: Quality pe through positive sport experiences.* Champaign, IL: Human Kinetics.
Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde geçerlilik ve güvenirlik*. Ankara: Seçkin.
Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics, 6th international edition (cover) edn*. New Jersey: Sage Publications, Thousand Oaks.
Taylı, A. (2006). *Akran yardımcılığı uygulaması aracılığıyla lise öğrencilerinde kişisel ve sosyal sorumluluğun arttırılması. (*Yayımlanmamış doktora tezi, Gazi Üniversitesi, Ankara). https://tez.yok.gov.tr/UlusalTezMerkezi/ adresinden edinilmiştir.
Taylı, A. (2013). Sorumluluğun bazı değişkenler açısından değerlendirilmesi. *Muğla Sıtkı Koçman Üniversitesi, Sosyal Bilimler Enstitüsü Dergisi, 30,* 68-84.
Trulson, M. E. (1986). Martial arts training as a "cure" for juvenile delinquency. *Human Relations*, *39*, 1131-1140.
Walsh, D. S. (2007). Supporting youth development outcomes: An evaluation of a responsibility model-based program. *Physical Educator*, *64*(1), 48-56.
Watson, D. L., Newton, M., & Kim, M. (2003). Recognition of values-based constucts in a summer physical activity program. *Urban Review*, *35*, 217-232.
Wright, Y. L. (2011). Relating recycling: Demographics, attitudes, knowledge and recycling behavior among UC Berkeley students. *UC Berkeley Student Recyling*. Retrieved March 25, 2017 from https://nature.berkeley.edu/ classes/es196/projects/2011final/WrightY_2011.pdf
Wright, P.M., & Burton, S. (2008). Implementation and outcomes of a responsibility-based physical activity program integrated into an intact high school physical education class. *Journal of Teaching in Physical Education*, 27(2), 138-154.
Yalom, I. (2001). *Varoluşçu psikoterapi.* (Çev. Z. İ. Babayiğit). (3. bs.). İstanbul: Kabalcı.
Yavuzer, H. (2006). *Çocuk eğitimi el kitabı*. (20. bs.). İstanbul: Remzi Kitabevi.

GENİŞ ÖZET

Giriş

Büyük ilgi çekmiş bir program olan Don Hellison'un (2011) bireysel ve sosyal sorumluluk modeli (BSSM), başlangıçta risk grubundaki gençler için tasarlanmıştır. Sorumluluk modeli olarak da bilinir. Modelde uygulanan programın sorumluluğu uygulayıcıdan katılımcılara doğru yön değiştirir (Hellison ve Martinek, 2006; Hellison, 2011; Hellison ve Walsh, 2002). Hellison ve Martinek (2006), bu modelin kapsayıcı amacını, kendi refahları ve başkalarının refahlarına katkıda bulunmak için

çocukların sorumluluk gelişimlerine yardımcı olmak olarak belirtir. BSSM, beden eğitimi sınıfları, danışmanlık ve antrenörlük programları, genişletilmiş günlük programlar gibi çeşitli alanlarda hizmet vermektedir ve programlar bu farklı türleri karşılamak için esnektir (Hellison, 2011; Lee ve Martinek, 2009; Martinek, Schilling ve Johnson, 2001; Walsh, 2007; Wright ve Burton, 2008).

Öncelikle ilgili literatüre göre, ülkemizde beden eğitimi ve spor alanında sorumluluk davranışları ile ilgili bir ölçeğe rastlanılmış olması, dünyada ise çok az sayıda olması sebebiyle; ayrıca, sorumluluk davranışlarının değerlendirilmesinde öğretmenlerin puanlama için harcadıkları zamanı kısaltması, öğrenci çalışmalarının değerlendirmesini basitleştirmesi, ortaya gözle görülür bir ölçüt çıkması ve soyut bir kavram olan sorumluluk davranışlarını kazandırmada eğitimin etkisine ilişkin geri bildirim alabilmek açılarından bu ölçeğin geliştirilmesi uygun görülmüştür. Bu bağlamda çalışmanın amacı, beden eğitimi ve spor alanında öğrencilerin sorumluluk davranışlarının belirlenmesi amacıyla bireysel ve sosyal sorumluluk modeline uygun bir ölçek geliştirmek; ve ölçek kullanılarak cinsiyet, yaş ve spor yapma yılı değişkenleri açısından öğrencilerin sorumluluk düzeylerini incelemektir.

Yöntem

Çalışma Grubu

Çalışma grubunu, AFA için Ankara ili Yenimahalle ilçe merkezinde, 2015-2016 yılı eğitim ve öğretim faaliyetlerini sürdüren Milli Eğitim Bakanlığı'na (MEB) bağlı liselerde, 128'i (% 38,8) kız, 202'si (% 61,2) erkek toplam 330 lise öğrencisi oluşturmuştur. DFA için ise 93'ü (% 37,2) kız, 157'si (% 62,8) erkek toplam farklı 250 lise öğrencisi oluşturmuştur

Veri Toplama Aracı

Çalışmada araştırmacı tarafından geliştirilen Bireysel ve Sosyal Sorumluluk Davranışları Ölçeği (BSSD-Ö) kullanılmıştır. Ölçek maddeleri, beden eğitimi ve spor alanında sorumluluk davranışlarını kazandırmada en uygun model olduğu düşünülen Hellison'un (2011) BSSM'nden ve Li ve ark. (2008)'nın geliştirdikleri ölçek maddelerinden faydalanılarak hazırlanmıştır (1, 2, 3 ve 8. maddeler).

Veri Analizi

Verilerin analizinde SPSS 20.0 paket programı ve LİSREL 8.80 kullanılmıştır. Ölçeğin yapı geçerliği için öncelikle ölçek puanlarının homojenliği kontrol edilmiş, sonrasında Açımlayıcı Faktör Analizi (AFA) ve Doğrulayıcı Faktör Analizi (DFA) yapılmıştır (Büyüköztürk, 2010). Madde geçerliliğine kanıt sağlamak amacıyla ölçeğin alt boyutlarında bulunan maddelerin birbirleri ile olan korelasyonları incelenmiştir. Ölçeğin iç tutarlılık güvenirliği için Cronbach Alpha katsayısı hesaplanmıştır. İki alt faktör arasındaki doğrusal ilişkiyi test etmek için Pearson Korelasyon analizi yapılmıştır. Bağımsız değişkenlere göre öğrencilerin ölçekten aldıkları puanların karşılaştırılması için öncelikle ölçek puanlarının homojenliği kontrol edilmiş, sonrasında ikili gruplarda bağımsız gruplar için t testi, ikiden fazla grupların ortalamaları arasındaki farkı test etmek için tek yönlü Varyans analizi kullanılmıştır (Büyüköztürk, 2010; Field, 2005).

Bulgular

Ölçeğin yapı geçerliğini belirlemek amacıyla 330 lise öğrencisi üzerinde yapılan AFA sonucunda iki alt faktörlü bir yapı elde edilmiştir. Elde edilen yapının uyum iyiliği değerlerine ulaşmak için farklı 250 lise öğrencisi üzerinde yapılan DFA sonucunda, ölçeğin faktör yapısı toplanan verilerle kabul edilebilir uyum göstermiştir.

Ölçeğin güvenirliğini tespit etmek için hesaplanan iç tutarlılık katsayılarında; iki faktöre ilişkin ∞=.82, birinci alt faktöre ilişkin ∞=.78, ikinci alt faktöre ilişkin ∞=.70 olarak bulunmuştur. İki faktör arasındaki korelasyon incelendiğinde faktörler arasında anlamlı ilişki olduğu görülmüştür.

İki alt faktördeki maddelerin ortalamaları alınarak yapılan %95'lik güvenirlik Pearson Korelasyon analizi sonucunda bireysel sorumluluk ve sosyal sorumluluk maddeleri arasında pozitif yönlü anlamlı doğrusal bir ilişki olduğu belirlenmiştir [$r(330)=.516$; $p<.01$].

Yapılan analizler sonucunda; öğrencilerin cinsiyet ve yaş değişkenine göre sorumluluk davranışlarına yönelik yargılara verilen cevapların puan ortalamaları arasında anlamlı bir farklılık bulunmamış ($p>.05$), spor yapma yılı değişkenine göre sosyal sorumluluk [$F(3,326)= 2.647$, $p<.05$] alt boyutunu algılayışlarına yönelik yargılara verdikleri cevapların puan ortalamaları arasında anlamlı bir farklılık bulunmuş; bireysel sorumluluk [$F(3,326)= .752$, $p>.05$] alt boyutunda ise bulunmamıştır.

Sonuç

Bu çalışmada ilk olarak, bireysel ve sosyal sorumluluk modeline uygun olarak öğrencilerin ve sporcuların sorumluluk davranışlarını değerlendirebilmek amacıyla "Bireysel ve Sosyal Sorumluluk Davranışları Ölçeği" geliştirilmiştir. Ölçek maddeleri kavramsal olarak BSSM literatürü ile uyumludur. Analiz sonucunda ölçeğin bireysel ve sosyal sorumluluk yapıları arasında pozitif yönlü bir ilişkili olduğu tespit edilmiştir.

Bu çalışmada ikinci olarak, geliştirilen BSSD-Ö aracılığıyla yaş, cinsiyet ve spor yapma yılı değişkenlerine göre sporcu öğrencilerin bireysel ve sosyal sorumluluk davranışları incelenmiştir. Bu çalışma, spor yapma yılına göre sporcu öğrencilerin sosyal sorumluluk davranışlarında anlamlı farklılıklar olduğunu ortaya koymuştur. Elde edilen sonuca göre, erken yaşlarda başlayarak uzun süre spor yapmanın, sporcu öğrencilerin sosyal sorumluluk davranışlarını olumsuz yönde etkilediği söylenebilir.

Araştırma sonuçlarına bakıldığında, lise çağındaki sporcu öğrencilerin bireysel ve sosyal sorumluluk davranışlarının cinsiyet ve yaşa göre farklılık göstermediği, erken yaşta başlayarak uzun süre spor yapan öğrencilerde ise sosyal sorumluluk davranışlarını sergilemede azalma olduğu belirlenmiştir. Bu sonuçlara göre; ilkokul, ortaokul, lise ve üniversite öğrencileri, hatta yetişkinler, ayrı gruplandırılarak farklı popülasyonlar arasında bireylerin bireysel ve sosyal sorumluluk davranışları karşılaştırılabilir. Farklı değişkenlerle bireylerin sorumluluk davranışları incelenebilir. Bireylere farklı yaş gruplarında sorumluluk eğitimi verilerek yaş grupları arasındaki sorumluluk davranışlarının değişimi incelenebilir. Spor yılının artması ile sosyal sorumluluk davranışlarının düşmesinin nedenleri araştırılabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);17-32

Madde Tepki Modellemesinde Genellenebilirlik İle İki Yüzeyli Desenlerin İncelenmesi*

Investigation of Two Facets Design With Generalizability In Item Response Modeling

Gülden KAYA UYANIK** Selahattin GELBAL***

Öz

Bu çalışmada, Madde Tepki Modellemesinde Genellenebilirlik (MTMG) yaklaşımı iki yüzeyli bx(m:t) deseni ile incelenmiş ve Genellenebilirlik Kuramından (GK) elde edilen sonuçlar ile karşılaştırılmıştır. Çalışmada simülasyon verisi kullanılmıştır. Genellenebilirlik Kuramı doğrusal veri seti bx (m:t) dengelenmiş rastgele deseni için üretilmiştir. Üretilen veriler madde takımı etkisi, madde takımı uzunluğu ve madde takımı sayısı açısından farklılık göstermektedir. Veriler toplamda iki evrenden ve her evren dört farklı koşuldan oluşmaktadır. Araştırmanın sonucu tüm evrenlere ait koşulların varyans kestirimlerinin MTMG yaklaşımı ve GK ile elde edilen sonuçlar arasında bir fark olmadığını göstermektedir. Elde edilen bu sonuç MTMG yaklaşımını ortaya atan ve tek yüzeyli desen üzerinde inceleyen Briggs ve Wilson'ın yapmış oldukları çalışma ile desteklenmektedir. MTMG yaklaşımı ve GK ile kestirilen değerler arasında fark yoktur; ancak MTMG yaklaşımında hata varyansı etkileşim varyansından ayrı olarak gözlenebilir. Çalışmada ayrıca madde takımları güvenirliği farklı koşullar altında incelenmiştir. Birey-madde takımı etkileşiminin küçük olduğu durumlarda etkileşimin büyük olduğu durumlara göre daha yüksek güvenirlik elde edilmiştir. Bunun yanında madde takımı etkisi arttıkça güvenirliğin düştüğü gözlenmiştir. Ayrıca tüm evrenlere ait koşullar incelendiğinde madde takımları için madde sayısı artıkça güvenirliğin arttığı gözlenmiştir.

Anahtar Kelimeler: Madde tepki modellemesinde genellenebilirlik, genellenebilirlik kuramı, madde takımı, madde sayısı, güvenirlik

Abstract

An approach called generalizability in item response modeling (GIRM) is investigated with two facets sx(i:t) design and results are compared with results of generalizability theory in this study. In this study simulated data is used. In Generalizability Theory linear model random facets balanced bx(m:h) design are used for generating data. Generated data are differed by factors. These factors are testlet effect, testlet length and number of testlet. All generated data consist of two different universes and all universes have four different conditions. According to the results of this study the estimates of variance components obtained using GIRM approach are generally quite similar to those obtained using GT approach. Briggs and Wilson's study is supported this result. There is no difference between results of GIRM and GT but error variance could be separated from residual variance with GIRM. This study also examines the reliability of testlets under different conditions. Testlets are more reliable when person-item variance is smaller. Furthermore, when testlet effect is increased,reliability is decreased. When conditions of all universes are investigated it is concluded that it is effective to have more items to increase reliability.

Keywords: Generalizability in item response modeling, generalizability theory, testlet, number of item, reliability

* Bu çalışma Gülden KAYA UYANIK'a ait doktora tezinden üretilmiştir.

** Doktor Öğretim Üyesi, Sakarya Üniversitesi, Eğitim Fakültesi, Sakarya-Türkiye, guldenk@sakarya.edu.tr. ORCID ID: orcid.org/0000-0002-8100-6994
*** Prof. Dr. , Hacettepe Üniversitesi, Eğitim Fakültesi, Ankara-Türkiye, sgelbal@gmail.com. ORCID ID: orcid.org/0000-0001-5181-7262

Geliş Tarihi:07.11.2017
DOI: 10.21031/epod.349718 *Kabul Tarihi:01.03.2018*

GİRİŞ

Eğitim çalışmalarında Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK) olmak üzere üç temel kuram yer almaktadır. Bazı araştırmacılar test kuramlarını puanların analiz edilmesi ve yorumlanması bakımından klasik ve modern olmak üzere ikiye ayırırlar. Klasik Test Kuramı ve Genellenebilirlik Kuramı klasik; Madde Tepki Kuramı ise modern kuram olarak ele alınmakta olup, bu kuramlarda farklı matematiksel modeller kullanılmaktadır. Modern test kuramının popülerliği gün geçtikçe artsa da KTK hâlâ en pratik kuram olarak görülmektedir. Günümüzde hala birçok test KTK'ya göre geliştirilmektedir. Genellenebilirlik Kuramının hata kaynaklarını çözümlemede ANOVA'yı kullanıyor olması GK'nın KTK'nın uzantısı olduğunu göstermektedir (Feldt ve Quails,1989; Shavelson ve Webb, 1991). Bu nedenle GK'da sıklıkla MTK'dan farklı olarak klasik kuramın içinde ele alınmaktadır.

Modern ve klasik kuramların tamamen farklı olmadığı, bir arada ya da birbirinin tamamlayıcısı olarak kullanılabileceğine yönelik iddialar da söz konusudur. Buna dayalı olarak araştırmacılar KTK ve GK'yı, MTK ile birbirine bağlayan çalışmalar üzerine yoğunlaşmışlardır. Örneğin Kolen ve Harris (1987) hem MTK hem de KTK'ya dayalı olarak çok değişkenli test modelleri ortaya atmışlardır. Benzer şekilde güvenirlik konusu içinde MTK'yı klasik kuramlarla birleştiren çalışmalar da söz konusudur. Samejima (1977,1994). güvenirlik ve ölçmenin standart hatası tahmini için KTK ile MTK'yı birbirine test bilgi fonksiyonu üzerinden bağlamıştır ve 1994 yılında yaptığı çalışma ile test bilgi fonksiyonu için tahmini güvenirlik önermiştir. Lord (1983), paralel formların yeteneğe dayalı güvenirlik katsayılarının kestirimi için eşitlikler ileri sürmüştür. Raju ve Oshime (2005), kısa ve uzun testler için yeteneğe dayalı güvenirlik kestirimini yapan iki eşitlik ortaya koymuştur ve bu eşitliklerden birinin Spearman Brown eşitliği ile aynı olduğunu ispatlamıştır. Dimitrov (2003), ikili puanlanan maddeler için gerçek puan kestirimlerini, MTK ve KTK'yı birleştirerek elde edilen eşitlikler üzerine çalışmıştır.

Madde Tepki Kuramında bireylerin belli bir alanda doğrudan gözlenemeyen yetenekleri ile bu alanı yoklayan sorulardan oluşan test maddelerine verdikleri yanıtlar arasındaki matematiksel ilişki yer alırken, Genellenebilirlik Kuramında ölçme sonuçlarının güvenirliği belirlenir, güvenilir gözlemler tasarlanır, araştırılır ve kavramsallaştırılır. Madde Tepki Kuramı (MTK) ve Genellenebilirlik Kuramı (GK) en azından yüzeysel açıdan birbirinden farklı olarak görülür. Örneğin; Brennan (2001), Genellenebilirlik Kuramının örnekleme modeli, Madde Tepki Kuramının ise bir ölçekleme modeli olduğunu belirtmiştir. Ancak her iki yaklaşım da desene ve ölçme araçlarının analizine ilişkin önemli bilgiler sağladığı için MTK'yı ve GK'yı hem büyük ölçekli sınavlarda hem de daha küçük ölçek çalışmalarında beraber kullanmak yararlı olabilir (Bock, Brennan ve Muraki, 2002). GK ve MTK'nın beraber kullanılmasının önemi anlaşılmış olmasına rağmen birleşimi oluşturmak zaman almıştır. Konuyla ilgili ilk adım Linacre (1989, 1999) tarafından atılmıştır. Linacre puanlayıcılar tarafından ikili puanlanan madde puanlarını incelemiştir. Elde ettiği model Rasch modelin (Rasch, 1960) basit bir genellemesi olarak sunulmuştur.

Alanyazında GK ile MTK'nın birlikte kullanıldığı çalışmalarda genel olarak her maddenin birden çok puanlayıcı tarafından puanlandığı desenler üzerinde çalışılmıştır (Alkahtani, 2012; Bock, Brennan ve Muraki, 2002; Kim ve Wilson, 2008; Patz, Junker, Johnson ve Mariano, 2002; Verhelst ve Verstralen, 2001; Wilson ve Hoskens, 2001; Zhang ve Roberts, 2013). Ancak yapılan bu çalışmalar MTK'nın yerel bağımsızlık varsayımını ihlal ettiği gerekçesiyle eleştirilmiştir. Glas (1989) farklı puanlayıcıların verdikleri puanların öğrencilerin cevaplarına bağlı olduğunu; bu nedenle bu modelin MTK'nın yerel bağımsızlık varsayımını ihlal ettiğini öne sürmüştür. Bu durumla başa çıkmak için MTK'da başka modeller üzerinde çalışılmıştır. Örneğin ilk olarak Zwinderman (1991) MTK modeli ile yapısal ANOVA modelini birleştirme üzerine çalışmalar yapmıştır. Daha sonra Fox ve Glas (2001) çalışmaların üzerinde durmuş; ancak kesin sonuca ulaşamamıştır. MTK ile GK'nın birleştirilmesi ilk olarak Briggs ve Wilson'ın (2004, 2007) yapmış oldukları çalışmalar ile gerçekleşmiştir.

Briggs ve Wilson (2004, 2007) çalışmalarında örnekleme modeli olan Genellenebilirlik Kuramını ölçekleme modeli olan MTK ile birleştirmiş ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) isimli yeni bir model ortaya atmışlardır. Yaptıkları çalışmada MTMG yaklaşımının GK'yı MTK'nın içine ilgili madde yüzeyinde dağılımsal varsayımlar yaparak dâhil ettiklerini ileri

sürmüşlerdir. MTMG yaklaşımında, Genellenebilirlik Kuramında geleneksel olarak kullanılan gözlenen puanlar yerine Markov Zinciri Monte Carlo (MZMC) tekniği ile elde edilen beklenen puanlar kullanılmıştır. MTMG'de MZMC yönteminin esnekliğinden de yararlanılarak GK varyans bileşenlerini MTK parametreleri üzerinden kestirmek mümkündür.

Briggs ve Wilson (2004, 2007) çalışmalarında GK ve MTMG arasındaki farkları şu şekilde sıralamışlardır:

- GK varyans bileşenleri kestiriminde gözlenen puanları kullanırken MTMG beklenen puanlar matrisini kullanır.
- Etkileşim varyansı ve hata GK'da birbirinden ayrılamazken MTMG yaklaşımında ayrı ayrı kestirilebilir.

MTMG yaklaşımında beklenen puanlar matrisinin kullanılması daha güvenilir kestirimler yapmaya olanak sağlamaktadır. Diğer yandan GK'nın en büyük dezavantajlarından biri olan hata varyansı sorunu MTMG yaklaşımı ile çözümlenebilir hale gelmektedir.

MTMG yaklaşımının tanıtılması ve örneklendirilmesi bireylerin maddeler ile çaprazlandığı " bxm" deseni üzerinden yapılmıştır. Çalışmada ölçme modeli olarak MTK'nın Rasch modeli, yapısal model olarak GK kullanılmıştır. MTMG yaklaşımı Briggs ve Wilson'ın (2004,2007) yapmış oldukları çalışma ile sınırlı kalmıştır. Yaklaşımın farklı desenler ve farklı çalışma koşulları altında nasıl sonuçlar verdiği henüz bilinmemektedir.

GK ve MTK'nın kullanıldığı diğer çalışma konusu ise madde takımlarıdır. Madde takımları; sınavı alan bireylerin takip edeceği, önceden belirlenmiş belli sayıdaki yolu içeren tek bir konuya ait bir grup ilişkili madde olarak tanımlanır (Wainer, Lewis; 1990). Madde takımları ile kestirilen yetenek; hem bireyin genel yeteneği hem de konu ile ilgili belirli bir yeteneğe bağlıdır (Demars, 2006). Tek uygulama ile birden çok cevap bulan madde takımlarının kullanımı özellikle zamansal açıdan ekonomik olması nedeniyle son zamanlarda artmıştır.

Madde takımları konusu hem GK ile hem de MTK ile ele alınabilir. GK'ya göre incelenen madde takımlarında; madde takımları maddelerin içinde yuvalandığı bir desen olarak ele alınmalıdır. Madde takımları yok sayılıp bir yüzey olarak ele alınmadığında, ölçmenin standart hatası için alt; güvenirlik için ise üst kestirim yapılır (Yen, 1993; Wainer ve Thissen, 1996; Wainer ve Wang, 2000).

Madde takımları maddelerin sahip olduğu karakteristik özelliklerden dolayı Madde Tepki Kuramında yer alan yerel bağımsızlık varsayımını ihlal eder. Yapılan birçok araştırmada madde takımlarının yerel bağımsızlığı bozduğu sonucuna ulaşılmıştır (Rosenbaum, 1988; Yen, 1993; Wainer, 1995; Wainer ve Thissen, 1996; Jiao, Kamata, Wang ve Jin, 2012). Yerel bağımsızlık varsayımının karşılanmadığı durumlarda bireylerin performansları, madde parametreleri ya da test istatistikleri için elde edilen sonuçlar hatalıdır (Yen, 1993; Wainer, 1995; Wainer ve Thissen, 1996; Ferrara, Huynh ve Bagli, 1997; Ferrara, Huynh ve Michaels, 1999; Bradlow, Wainer ve Wang 1999). Madde takımı puanlamada, madde takımını oluşturan maddelerin birbirlerine bağlı olmalarının göz önünde bulundurulması olumlu bir durum iken; bu maddeleri cevaplandıran bireyin cevap deseni ile ilgili bilgi kaybı söz konusudur. Bu olumsuzluğu ortadan kaldırmak adına orijinal MTK modellerine kişiye özgü tesadüfî madde takımı etkisinin de eklenmesi yeni bir strateji olarak ele alınmaktadır (Bradlow, Wainer ve Wang, 1999; Wainer, Bradlow ve Du, 2000; Wang, Bradlow ve Wainer, 2002; Li, Bolt ve Fu, 2006). Bu strateji ise "Madde Takımı Tepki Kuramı" (MTTK) olarak adlandırılmaktadır (Wainer, Bradlow ve Du 2000). Dresher (2004) MTTK modelinin kullanılmasının, madde ayırıcılık ve güçlük parametrelerinin madde takımını oluşturan maddelerin birbirlerine bağımlılıklarını yok sayan madde takımı puanlama ya da tek boyutlu MTK modellerine göre daha iyi kestirim yaptığını bulmuştur (Akt. Chien, 2008). MTTK'da temel olarak kişiye özgü tesadüfî madde takımı etkisini ele alan pek çok madde takımı modeli bulunmaktadır (Bradlow, Wainer ve Wang, 1999; Wainer, Bradlow ve Du, 2000; Wang, Bradlow ve Wainer, 2002; Li, Bolt ve Fu, 2006). Bütün MTTK modelleri, her bir bireydeki yerel madde bağımlılığı miktarını belirtmek için, geleneksel MTK parametrelerinin yanında bir de madde takımı parametresini önermektedir. Genel olarak, geliştirilen bütün MTTK modelleri, çok boyutlu MTK modellerinden ya da daha önce önerilen bir MTTK modelinden uyarlanmıştır.

Briggs ve Wilson'ın (2004,2007) yaptıkları çalışmalar ile ortaya atılan MTMG yaklaşımı birbirinden farklı olan MTK ve GK modelini birbirine bağlar. Bu bağlam bir bireyin tek bir madde için beklenen puanı üzerinden yapılır. Bu durumun, tek yüzeyli desenler için mümkün olduğu Briggs ve Wilson'ın (2004, 2007) yaptığı çalışmalar ile ispatlanmıştır.

MTMG yaklaşımı elde edilen ümit verici sonuçların yanında alanyazında çok az gerçekleştirilen birbirlerinden farklı iki ölçme kuramını bir arada kullanması açısından değerlidir. Ancak MTMG çalışmaları tek yüzeyli desenler ile sınırlı kalmıştır. MTMG yaklaşımının çok yüzeyli durumlarda nasıl işlediğini bilmek gereklidir. Örneğin MTMG çok yüzeyli modellerde çalışmazsa adres gösterilen sorunlarda GK'nın yerine kullanılabileceği söylenemez. Bu nedenle bu çalışmanın temel amacı Briggs ve Wilson'ın (2004, 2007) yaptığı çalışmayı tek yüzeyli desenden çok yüzeyli desene çıkartmaktır. Çalışmada kullanılan çok yüzeyli desen, maddelerin (m) takımlara (t) yuvalandığı ve bireylerin (b) bunlarla çaprazlandığı rastgele dengelenmiş yuvalanmış desendir. Bu desen simgesel olarak bx(m:t) olarak gösterilir.

Madde takımlarından oluşan testlerin kullanımı, beraberinde getirdiği avantajlar nedeniyle hem ulusal hem de uluslararası sınavlarda artış göstermektedir. Ancak madde takımlarının yerel madde bağımsızlığı ihlali göz ardı edildiğinde kestirimde hatalara neden olduğu açıktır (Dresher,2004). Bu çalışmada madde takımlarının farklı koşullar altında elde edilen parametreleri incelenmiştir. Bu koşullar; birey-madde takımlarının etkileşimlerinin varyans büyüklükleri, madde takımları sayısı ve madde takımlarında bulunan madde sayısıdır.

Bu çalışmada birbirinden farklı kuramlarının birleştirmesi amaçlanmıştır. Aynı zamanda madde takımının olduğu farklı durumlar, farklı kuramlar çerçevesinde incelenmiştir. Bu nedenle, hem farklı kuramların bir arada kullanılması hem de madde takımlarının sıklıkla kullanıldığı ulusal çapta yürütülen geniş ölçekli sınavlarda parametre kestirimleri için farklı bir yaklaşım önermesi açısından önemli olduğu düşünülmektedir.

YÖNTEM

Bu araştırmada farklı koşullar için Madde Tepki Modellemesinde Genellenebilirlik yaklaşımı (MTMG) ile sonuçlar elde edilmiş ve elde edilen sonuçlar aynı koşullar için Genellenebilirlik Kuramı (GK) ile elde edilen sonuçlar ile karşılaştırılmıştır. Kontrollü koşulların oluşturularak uygun verilerin türetilmesi bakımından araştırma, bir simülasyon çalışmasıdır. Araştırmada simülasyon verileri ile farklı koşullar oluşturulmakta ve koşulların durumları/sonuçları değerlendirilmektedir. Araştırma bu yönüyle de yöntemlerin geliştirilmesine katkı sağlayacağından temel araştırma olarak kabul edilebilir (Karasar, 2004).

Çalışma Verileri

Çalışmada simülasyon veri kullanılmıştır. Verilerin üretimi için 100 tekrar yapılmış böylelikle hata en aza düşürülmeye çalışılmıştır. İki farklı evren her evrene ait dört farklı koşuldan oluşan 8 farklı veri seti vardır. Her veri seti için simülatif olarak üretilen veriler 10 tekrar üzerinden yürütülmüştür. Ancak bulgular kısmında tüm tekrarların ortalaması paylaşılmıştır.

GK doğrusal veri seti bx (m:t) dengelenmiş rastgele deseni için üretilmiştir. Oluşturulan tüm veri setlerinde birey sayısı ve 1-0 şeklinde puanlanan toplam madde sayısından oluşan n_b x n_m gözlenen puan matrisi elde edilmiştir. Toplam madde sayısı; madde takımı sayısı ve madde takımlarında yer alan maddelerin çarpımı ile elde edilmiştir. Her madde sadece bir madde takımında yuvalanmıştır. n_m ve n_t farklı çalışma koşullarına göre değişiklik gösterir ancak birey sayısı (n_b) bu çalışma için 1000'e sabitlenmiştir.Oluşturulan veriler madde takımı etkisi, madde takımı uzunluğu ve madde takımı sayısı açısından farklılık göstermektedir.

Madde Takımı Etkisi

Veri üretmek için uyarlanan Genellenebilirlik Kuramı doğrusal veri seti bx(m:t) dengelenmiş rastgele deseni için üretilmiştir. GK doğrusal veri setinde madde takımı etkisi olan $\sigma^2_{(t)}$ yerine madde takımlarının anlam ya da zorluk bakımından bireyden bireye farklılık gösterip göstermediğini belirten birey ve madde takımı etkileşiminin ($\sigma^2_{(bt)}$) iki farklı durumu kullanılmıştır. Bu durumlar birey-madde takımı etkileşiminin diğer varyans kaynakları arasında en büyük değere ve en küçük değere sahip olması bakımından değişiklik gösterir. Varyans kaynaklarının değerleri gerçek durumları yansıtması için yapılan birçok çalışma (Lee ve Frisbie, 1999; Lee, Brennan ve Frisbie, 2000; Chien 2008) incelenerek oluşturulmuştur.

Madde Takımı Uzunluğu ve Madde Takımı Sayısı

Madde takımı uzunluğu kullanılan madde takımında bulunan madde sayısını; madde takımı sayısı ise testte bulunan toplam madde takımı sayısını ifade etmektedir. Bu çalışmada bx(m:t) deseni dengelenmiş olarak inceleneceği için madde takımı uzunluğu ve madde takımı sayısı çarpımı toplam madde sayısını vermektedir. Çalışmada gerçek durumlara uygun olması açısından uluslararası (PIRLS; ITBS RC) ve ulusal sınavlar (ALES, KPSS, bazı üniversitelere ait hazırlık geçme sınavları) incelenmiş ve madde takımı uzunluklar 6 ve 9; madde takımı sayıları 3 ve 5 olarak belirlenmiştir. Belirlenen madde takımı uzunluğu ve madde takımı sayılarına göre veride kullanılan madde sayıları sırasıyla 18, 30, 27, 45'tir. Aşağıda yer alan Tablo 1'de çalışma koşulları özetlenmiştir.

Tablo 1. Araştırmada Yer Alan Çalışma Koşulları

	GK doğrusal veri seti
Madde takımı etkisi	Küçük σ^2(bt)
	Büyük σ^2(bt)
Madde takımı uzunluğu - madde takımı sayısı	6-3
	6-5
	9-3
	9-5
Özet Çalışma koşulları	A evreni
	Küçük ve 6-3
	Küçük ve 6-5
	Küçük ve 9-3
	Küçük ve 9-5
	B evreni
	Büyük ve 6-3
	Büyük ve 6-5
	Büyük ve 9-3
	Büyük ve 9-5

Verilerin Analizi

Araştırmanın verileri R programı ile üretilmiştir. A ve B olarak isimlendirilen her evrene ait 4 koşul bulunmaktadır. Toplamda 8 farklı durumunun her biri için 10 farklı veri elde edilmiştir. Üretilen verilerin Madde Tepki Modellemesinde Genellenebilirlik çözümlemeleri WinBUGS, Genellenebilirlik Kuramı çözümlemeleri ise EDUG programı ile yapılmıştır ve her durum için bu kestirimler karşılaştırılmıştır.

BULGULAR

Araştırma problemine ait bulgulardan önce simülasyon verisine ait test güvenirlikleri ve madde güçlükleri ile ilgili bilgilere yer verilmiştir. Tablo 2'de iki evrenden elde edilmiş veri setinin farklı koşullarına ait 10 tekrardan oluşan verilerin ortalama güvenirlik değeri, standart sapma değerleri ve

verilerin almış olduğu maksimum ve minimum güvenirlik değerleri yer almaktadır. Veri setlerine ait güvenirlik katsayıları genellenebilirlik katsayılarıdır ve ölçme objesi olarak bireyler alınmıştır.

Tablo 2. Veri Setlerine Ait Güvenirlik Değerleri

	Evren	*Koşullar*	*SS*	*Max. güvenirlik*	*Min. güvenirlik*	*Ortalama güvenirlik değeri*
GK doğrusal veri seti	A	Küçük ve 6-3	0,011	0,648	0,609	0,624
		Küçük ve 6-5	0,016	0,756	0,699	0,734
		Küçük ve 9-3	0,015	0,723	0,678	0,698
		Küçük ve 9-5	0,009	0,821	0,779	0,800
	B	Büyük ve 6-3	0,018	0,607	0,552	0,579
		Büyük ve 6-5	0,023	0,723	0,659	0,691
		Büyük ve 9-3	0,014	0,664	0,625	0,653
		Büyük ve 9-5	0,008	0,758	0,740	0,750

Tablo 2'de yer alan A ve B evrenlerine ait koşullardan elde edilen güvenirlik katsayıları incelendiğinde madde sayısı arttıkça güvenirliğin arttığı görülmektedir. Bunun yanında eşit madde sayıları için A evreninden kestirilen güvenirlik katsayıları B evrenine göre daha yüksektir. Bu beklenen bir sonuçtur çünkü B evreninde birey-madde takımı etkisi A evrenine göre daha büyük bir değerdedir ve bu değerin büyük olması hataya sebep olmaktadır.

Madde güçlüğü KTK'ya göre maddeye doğru cevap verenlerin yüzdesi; MTK'ya göre ise maddenin P olasılıkla doğru yanıtlanması için gerekli yetenek düzeyi olarak tanımlanır. Tablo 3'te farklı koşullar için elde edilen 10 tekrar verisine ait ortalama madde güçlüğü değerleri; standart sapma değerleri ve alınan maksimum ve minimum madde güçlüğü değerleri yer almaktadır.

Tablo 3. Veri Setlerine Ait Madde Güçlükleri

Evren	*Koşullar*	*SS*	*Max. değer*	*Min. değer*	*Ortalama madde güçlük değeri*
A	Küçük ve 6-3	0,110	0,920	0,385	0,650
	Küçük ve 6-5	0,095	0,915	0,390	0,680
	Küçük ve 9-3	0,105	0,880	0,355	0,650
	Küçük ve 9-5	0,085	0,890	0,355	0,675
B	Büyük ve 6-3	0,120	0,945	0,275	0,740
	Büyük ve 6-5	0,230	0,980	0,324	0,700
	Büyük ve 9-3	0,100	0,930	0,390	0,665
	Büyük ve 9-5	0,095	0,935	0,237	0,675

Tablo 3 incelendiğinde GK doğrusal veri seti için madde güçlüğü ortalamasının 0,6-0,7 aralığında olduğu görülmüştür. Veri setinde yer alan soruların kolay olmasının nedeni veri seti üretilirken kesme noktasının -0,2 seçilmiş olmasıdır. En zor soruya ait madde güçlüğü değeri 0,237 olduğu için çok zor bir sorunun olmadığı görülmektedir.

Simülasyon verisine ait test güvenirlikleri ve madde güçlükleri ile ilgili bilgilerin ardından birey-madde etkileşim varyansının diğer varyans değerleri arasında en küçük, madde takımı sayısı 3 veya 5 ve madde takımlarında yer alan madde sayısı 6 veya 9 olduğu durumlarda (A evreni) Madde Tepki Modelinde Genellenebilirlik yaklaşımına ve Genellenebilirlik Kuramına göre elde edilen; a) varyans bileşenleri arasında fark ve b) bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasındaki fark incelenmiştir.

Çalışmada varyans değerleri elde edilen değişkenlik kaynakları şunlardır;

Birey değişkenlik kaynağı ($\sigma^2(b)$): Değişkenliğin ilk kaynağı ölçme objesi olan öğrencilerin maddelerden aldıkları farklı puanlardır. Evren puanı için olan bu varyans bileşeni, bireylerin birbirinden ne derece sistematik bir şekilde farklılık gösterdiğini ifade etmektedir. Bu nedenle birey değişkenlik kaynağı değerinin olabildiğince büyük olması istenen bir durumdur.

Madde takımı değişkenlik kaynağı (σ^2(t)): Madde takımı değişkenliği madde takımları arasındaki tutarsızlıktan kaynaklanmaktadır. Diğer bir deyişle bir madde takımının bir bireye kolay gelirken diğer madde takımının aynı birey için zor gelmesi madde takımındaki varyansın sebebidir. Bu değerler, madde takımlarının birbiri arasındaki değişkenliğin derecesini vermektedir. Bu nedenle madde takımı varyans bileşeni değerinin olabildiğince küçük olması istenen bir durumdur.

Birey-madde takımı değişkenlik kaynağı (σ^2(bt)): Madde takımlarının bazı bireyler için kolaylık-zorluk anlamında tutarsızlıkları olabilir. Bu tutarsızlıkların derecesi birey- madde takımı değişkenlik kaynağında incelenir. Bu değerler, birey-madde takımlarının etkileşiminin derecesini vermektedir. Birey - madde takımı değişkenlik kaynağı A ve B evrenleri için manipüle edilen değişkenlik kaynağıdır. A evreninde bu değişkenlik kaynağı diğer değişkenlik kaynakları içerisinde en küçük B evrenine ise en büyük değer olarak elde edilmiştir.

Madde-madde takımı değişkenlik kaynağı σ^2(m:t): Her bir madde takımında yer alan maddelere ilişkin; bireyler bazı maddelerde geçmiş yaşantılarından dolayı daha avantajlı iken bazılarında bu durum söz konusu olmayabilir. Maddelerin güçlük düzeyleri arasındaki farklılıklar, maddeler farklı madde takımlarında yer aldığı için, her bir madde takımında yer alan maddeler üzerinden yorumlanıp, madde takımlarından bağımsız bir yorum yapılamaz. Bu değerler, maddelerin madde takımlarının içinde yuvalanmış olmasından kaynaklanan etki değerlerini verir.

Birey-madde-madde takımı (artık) değişkenlik kaynağı σ^2(bm:t): İki yüzeyli bx(m:t) deseninde yer alan son değişkenlik kaynağı birey, madde ve madde takımı etkileşiminin ve tesadüfî hataların yol açtığı değişkenliktir. GK'da etkileşim varyansı hata varyansından ayrılamazken MTMG'de hata varyansı ayrı olarak elde edilebilir. Bu nedenle birey- madde - madde takımı etkileşimi için elde edilen varyans değerleri için GK ve MTMG kestirimi arasındaki fark MTMG ile elde edilen hata varyansı değerini de kapsamaktadır.

Tablo 4'te A evrenine ait her bir değişkenlik kaynağı için Genellenebilirlik Kuramı (GK) ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) yaklaşımı ile kestirilen değerler ayrı ayrı verilmiş ve aralarındaki fark hesaplanmıştır.

Tablo 4. A Evrenine Ait Kestirilen Varyans Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
A evreni	σ^2***(b)***	***GK***	0,01982	0,02004	0,01988	0,02068
		MTMG	0,01980	0,02002	0,01987	0,02066
		FARK	0,00002	0,00002	0,00001	0,00002
	σ^2***(t)***	***GK***	0,00524	0,00527	0,00515	0,00488
		MTMG	0,00523	0,00526	0,00513	0,00487
		FARK	0,00001	0,00001	0,00002	0,00001
	σ^2***(bt)***	***GK***	0,00039	0,00112	0,00100	0,00111
		MTMG	0,00038	0,00111	0,00099	0,00110
		FARK	0,00001	0,00001	0,00001	0,00001
	σ^2***(m:t)***	***GK***	0,00920	0,01080	0,00995	0,01018
		MTMG	0,00919	0,01078	0,00994	0,01016
		FARK	0,00001	0,00002	0,00001	0,00002
	σ^2***(bm:t)***	***GK***	0,18642	0,18801	0,18787	0,19052
		MTMG	0,18582	0,18742	0,18728	0,18995
		FARK	0,00060	0,00059	0,00060	0,00057
	σ^2***(e)***	***MTMG***	0,00059	0,00056	0,00058	0,00055

A evreni için dört farklı koşul vardır. Birinci koşulda birey-madde takımı etkileşimin varyans değerinin diğer varyanslar arasında en küçük, madde takımında yer alan madde sayısı 6, madde takımı sayısı 3'tür. A evreninin ikinci koşulunda birey-madde takımı etkileşimi varyans değeri diğer varyanslar arasında en küçük, madde takımında yer alan madde sayısı 6 ve madde takımı sayısı 5'tir. Birey-madde takımı etkileşimin varyans değerinin diğer varyanslar arasında en küçük olduğu, madde takımında yer alan madde sayısı 9, madde takımı sayısı 3 olması durumu A evreninin üçüncü koşuludur. A evreninin dördüncü koşulunda birey-madde takımı etkileşimin varyans değeri diğer

varyanslar arasında en küçük, madde takımında yer alan madde sayısı 9, madde takımı sayısı 5'tir. Buna göre tüm koşullar için değişkenlik kaynaklarına ait değerlerin MTMG ve GK kestirimi arasındaki farklar incelendiğinde; İki yaklaşım ile elde edilen Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynaklarına ait varyans değerleri arasındaki fark her tekrar için 0,00002 ile 0,00001 arasında değişmektedir. Bu durum; A evreni tüm koşullarında Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynakları için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

Birey- madde - madde takımı etkileşimi varyans bileşeni için hesaplanan değer GK kestirimi ile tüm tekrarların ortalaması alındığında birinci koşulda GK kestirimi için bu değer 0,18642 MTMG kestirimi için ise 0,18582'dir. İkinci koşulda GK kestirimi için 0,18801 MTMG kestirimi için ise 0,18742'dir. Üçüncü koşulda GK kestirimi için bu değer 0,18787 MTMG kestirimi için ise 0,18728'dir. Dördüncü koşulda ise GK kestirimi için bu değer 0,19052 MTMG kestirimi için ise 0,18995'dir. Elde edilen bu değerler her iki yaklaşım içinde diğer varyans değerleri arasında birinci sırada yer almaktadır. Tablo 4'te yer alan σ^2(e) değerleri incelendiğinde MTMG yaklaşımı ile hesaplanabilen hata varyansının 0,00055 ile 0,00059 arasında değerler aldığı görülmektedir. MTMG ile elde edilen hata varyansı değerleri farktan çıkarıldığında birey- madde - madde takımı etkileşimi için kestirilen değerler arasındaki fark 0,00001 ile 0,00003 arasında olduğu görülür. Bu durum A evrenin her koşulunda birey-madde-madde takımı değişkenlik kaynağı için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

A evreninde birey ve madde takımı etkileşimi ($\sigma^2_{(bt)}$) diğer varyanslar arasında en küçük değerdir ve koşullar içinde madde takımı uzunluğu - madde takımı sayısı sırasıyla 6-3,6-5,9-3,9-5'tir. Yapılan analizler sonucunda A evreninin her koşulu için MTMG yaklaşımı ve GK yaklaşımı arasında bir fark bulunmamıştır.

A evrenine ait GK ve MTMG yöntemleri ile kestirilen bağıl ve mutlak hata varyans değerleri ile G ve Phi katsayılarını incelenmiş ve evrene ait koşullar çerçevesinde değerler ayrı ayrı verilmiş ve aralarındaki fark hesaplanmıştır. Tablo 5'te elde edilen değerler yer almaktadır.

Tablo 5. A Evreni Koşullarına Ait Bağıl ve Mutlak Hata Varyansı, Genellenebilirlik ve Phi Katsayısı Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
A evreni	***σ2(δ)***	***GK***	0,01208	0,00722	0,00856	0,00517
		MTMG	0,01207	0,00721	0,00854	0,00515
		FARK	0,00001	0,00001	0,00002	0,00002
	Eb²	***GK***	0,62410	0,73403	0,69841	0,80034
		MTMG	0,62408	0,73402	0,69840	0,80032
		FARK	0,00002	0,00001	0,00001	0,00002
	σ²(Δ)	***GK***	0,01292	0,00771	0,00900	0,00563
		MTMG	0,01291	0,00770	0,00899	0,00561
		FARK	0,00001	0,00001	0,00001	0,00002
	Φ	***GK***	0,60426	0,72102	0,68668	0,78625
		MTMG	0,60425	0,72100	0,68666	0,78623
		FARK	0,00001	0,00002	0,00002	0,00002

A evreninin her koşulu için Genellenebilirlik Kuramına göre kestirilen bağıl hata varyansı değerleri ile Madde Tepki Modellemesinde Genellenebilirlik yaklaşımına göre kestirilen bağıl hata varyansı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında her koşul altında bağıl hata varyansı kestiriminde fark olmadığı söylenebilir.

Bağıl hata varyansına bağlı olarak genellenebilirlik katsayısı değerleri de hesaplanmıştır. Genellenebilirlik katsayısı kestiriminde iki yaklaşım arasındaki fark tüm koşullarda her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında genellenebilirlik katsayısı kestiriminde fark olmadığı söylenebilir.

Benzer bir durum mutlak hata değerleri içinde geçerlidir. Mutlak hata varyansı kestiriminde iki yaklaşım arasındaki fark tüm koşullarda her tekrar için 0,00001 ile 0,00002 arasında değişmektedir. Bu nedenle iki yaklaşım arasında mutlak hata varyansı kestiriminde fark olmadığı söylenebilir.

Mutlak hata varyansına bağlı olarak hesaplanan Phi katsayısı değerleri için her koşul altında iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir, bu nedenle iki yaklaşım arasında Phi katsayısı kestiriminde fark olmadığı söylenebilir.

A evrenine ait koşullar madde takımı ve madde takımlarında yer alan madde sayıları açısından farklılık göstermektedir. Madde takımı uzunluğu 6 olan birinci ve ikinci koşul arasında madde takımı sayısı daha fazla olan ikinci koşuldan daha yüksek güvenirlik elde edilmiştir. Benzer şekilde madde takımı uzunluğu 9 olan üçüncü ve dördüncü koşullarda madde takımı sayısı fazla olan dördüncü koşuldan daha yüksek güvenirlik elde edilmiştir. Madde takım sayıları eşit olan birinci-üçüncü ve ikinci-dördüncü koşullarda madde takımı sayısı fazla olan koşullardan daha yüksek güvenirlik elde edilmiştir. Tüm bu bulgular her dört koşul incelendiğinde toplam madde sayısı arttıkça güvenirliğin arttığını göstermektedir.

Çalışmada ayrıca Birey-madde etkileşim varyansının diğer varyans değerleri arasında en büyük, madde takımı sayısı 3 veya 5 ve madde takımlarında yer alan madde sayısı 6 veya 9 olduğu durumlarda (B evreni) Madde Tepki Modelinde Genellenebilirlik yaklaşımına ve Genellenebilirlik Kuramına göre elde edilen; a) varyans bileşenleri arasında fark b) bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında fark incelenmiştir.

Tablo 6'da B evrenine ait her bir değişkenlik kaynağı için Genellenebilirlik Kuramı (GK) ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) yaklaşımı ile kestirilen değerler ayrı ayrı verilmiş ve aralarındaki fark hesaplanmıştır.

Tablo 6. B Evrenine Ait Kestirilen Varyans Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
B evreni	***σ2(b)***	***GK***	0,00956	0,00938	0,01002	0,00945
		MTMG	0,00954	0,00936	0,01000	0,00943
		FARK	0,00002	0,00002	0,00002	0,00002
	σ2(t)	***GK***	0,00517	0,00569	0,00515	0,00543
		MTMG	0,00515	0,00568	0,00513	0,00541
		FARK	0,00002	0,00001	0,00002	0,00002
	σ2(bt)	***GK***	0,01847	0,01784	0,01989	0,01836
		MTMG	0,01845	0,01782	0,01987	0,01834
		FARK	0,00002	0,00002	0,00002	0,00002
	σ2(m:t)	***GK***	0,00942	0,00603	0,00904	0,00894
		MTMG	0,00940	0,00602	0,00902	0,00892
		FARK	0,00002	0,00001	0,00002	0,00002
	σ2(bm:t)	***GK***	0,01808	0,01568	0,01848	0,01512
		MTMG	0,01751	0,01496	0,01732	0,01432
		FARK	0,00057	0,00072	0,00116	0,00080
	σ2(e)	***MTMG***	0,00055	0,00071	0,00115	0,00079

B evreninin birinci koşulunda birey madde takımı etkileşimi varyansı diğer varyans değerleri arasında en büyük, madde takımında yer alan madde sayısı 6 ve madde takımı sayısı 3'tür. B evreninin ikinci koşulunda birey madde takımı etkileşimi büyük, madde takımında yer alan madde sayısı 6 ve madde takımı sayısı 5'tir. B evreninin üçüncü koşulunda birey madde takımı etkileşimi büyük, madde takımında yer alan madde sayısı 9 ve madde takımı sayısı 3'tür. Dördüncü koşulda ise birey madde takımı etkileşimi büyük, madde takımında yer alan madde sayısı 9 ve madde takımı sayısı 5'tir.

Buna göre tüm koşullar için değişkenlik kaynaklarına ait değerlerin MTMG ve GK kestirimi arasındaki farklar incelendiğinde; İki yaklaşım ile elde edilen Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynaklarına ait varyans değerleri arasındaki fark her tekrar için 0,00002 ile 0,00001 arasında değişmektedir. Bu durum; B evreni tüm

koşullarında Birey-madde-madde takımı (artık) değişkenlik kaynağı (σ2(bm:t)) dışında kalan diğer değişkenlik kaynakları için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

Birey- madde - madde takımı etkileşimi varyans bileşeni için hesaplanan değer GK kestirimi ile tüm tekrarların ortalaması alındığında birinci koşulda GK kestirimi için bu değer 0,01808 MTMG kestirimi için ise 0,01751'dir. İkinci koşulda GK kestirimi için 0,01568 MTMG kestirimi için ise 0,01496'dir Üçüncü koşulda GK kestirimi için bu değer 0,01848 MTMG kestirimi için ise 0,01732'dir. Dördüncü koşulda ise GK kestirimi için bu değer 0,01512 MTMG kestirimi için ise 0,01432'dir. Elde edilen bu değerler her iki yaklaşım içinde diğer varyans değerleri arasında ikinci sırada yer almaktadır. Tablo 6'da yer alan σ2(e) değerleri incelendiğinde MTMG yaklaşımı ile hesaplanabilen hata varyansının 0,00055 ile 0,00115 arasında değerler aldığı görülmektedir. MTMG ile elde edilen hata varyansı değerleri farktan çıkarıldığında birey- madde - madde takımı etkileşimi için kestirilen değerler arasındaki fark 0,00001 ile 0,00002 arasında olduğu görülür. Bu durum B evrenin her koşulunda birey-madde-madde takımı değişkenlik kaynağı için GK ve MTMG kestirimi arasında fark olmadığı şeklinde yorumlanabilir.

B evreninde birey ve madde takımı etkileşimi ($\sigma^2_{(bt)}$) diğer varyanslar arasında en büyük değerdir ve koşullar içinde madde takımı uzunluğu - madde takımı sayısı sırasıyla 6-3,6-5,9-3,9-5'tir. Yapılan analizler sonucunda B evreninin her koşulu için MTMG yaklaşımı ve GK yaklaşımı arasında bir fark bulunmamıştır.

B evreninin tüm koşulları için Madde Tepki Modelinde Genellenebilirlik yaklaşımına ve Genellenebilirlik Kuramına göre elde edilen bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında fark olup olmadığı incelenmiştir.

Tablo 7'de B evreninin dört farklı koşuluna ait bağıl ve mutlak hata varyansları ile genellenebilirlik ve Phi katsayısı değerleri yer almaktadır.

Tablo 7. B Evreni Koşullarına Ait Bağıl ve Mutlak Hata Varyansı, Genellenebilirlik ve Phi Katsayısı Değerleri

			Birinci koşul	*İkinci koşul*	*Üçüncü koşul*	*Dördüncü koşul*
B evreni	***σ2(δ)***	***GK***	0,01361	0,00786	0,00934	0,00613
		MTMG	0,01359	0,00784	0,00933	0,00612
		FARK	0,00002	0,00002	0,00001	0,00001
	Eb^2	***GK***	0,57973	0,69163	0,65353	0,75004
		MTMG	0,57972	0,69162	0,65352	0,75002
		FARK	0,00001	0,00001	0,00001	0,00002
	$\sigma^2(\Delta)$	***GK***	0,01572	0,00974	0,01154	0,00713
		MTMG	0,01571	0,00973	0,01152	0,00711
		FARK	0,00001	0,00001	0,00002	0,00002
	Φ	***GK***	0,54355	0,63148	0,61845	0,72069
		MTMG	0,54354	0,63146	0,61843	0,72067
		FARK	0,00001	0,00002	0,00002	0,00002

B evreninin tüm koşulları için bağıl hata varyansı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında bağıl hata varyansı kestiriminde fark olmadığı söylenebilir.

Bağıl hata varyansına bağlı olarak genellenebilirlik katsayısı değerleri hesaplanmıştır. Genellenebilirlik katsayısı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir bu nedenle iki yaklaşım arasında genellenebilirlik katsayısı kestiriminde fark olmadığı söylenebilir.

B evreninin dört koşulu içinde hesaplanan mutlak hata varyansı değerleri incelendiğinde iki yaklaşım arasındaki farkın her tekrar için 0,00001 ile 0,00002 arasında değiştiği görülmektedir. Bu nedenle dört koşul için iki yaklaşım arasında mutlak hata varyansı kestiriminde fark olmadığı söylenebilir.

Mutlak hata varyansına bağlı olarak Phi katsayısı değerleri hesaplanmıştır. Her dört koşul için Phi katsayısı kestiriminde iki yaklaşım arasındaki fark her tekrar için 0,00001 ile 0,00002 arasında değişmektedir. Bu nedenle iki yaklaşım arasında Phi katsayısı kestiriminde fark olmadığı söylenebilir.

Bağıl hata varyansına bağlı olarak hesaplanan genellenebilirlik katsayısı değerleri ise GK kestirimi

B evrenine ait koşullar madde takımı ve madde takımlarında yer alan madde sayıları açısından farklılık göstermektedir. Madde takımında yer alan madde sayısı 6 olan birinci ve ikinci koşul arasında daha fazla madde bulunan ikinci koşuldan daha yüksek güvenirlik elde edilmiştir. Benzer şekilde madde takımında yer alan madde sayısı 9 olan üçüncü ve dördüncü koşullarda madde sayısı fazla olan dördüncü koşuldan daha yüksek güvenirlik elde edilmiştir. Madde takım sayıları eşit olan birinci-üçüncü ve ikinci-dördüncü koşullarda madde takımı sayısı fazla olan koşullardan daha yüksek güvenirlik elde edilmiştir. Tüm bu bulgular her dört koşul incelendiğinde toplam madde sayısı arttıkça güvenirliğin arttığını göstermektedir. A ve B evrenleri birey-madde takımı etkileşimi varyansı açısından farklılık göstermektedir. İki evren güvenirlik değerleri açısından karşılaştırıldığında her koşul için birey-madde takımı etkileşiminin küçük olduğu A evreninde daha yüksek güvenirlik elde edilmiştir.

SONUÇLAR ve TARTIŞMA

Araştırmada yer alan Genellenebilirlik Kuramı veri seti iki farklı evrenden (A ve B) ve her evren dört farklı koşuldan oluşmaktadır. Genellenebilirlik Kuramı (GK) ve Madde Tepki Modellemesinde Genellenebilirlik (MTMG) modeline göre A ve B evrenlerinin her koşulu için kestirilen birey ($\sigma^2(b)$), madde takımı ($\sigma^2(t)$), birey madde takımı etkileşimi ($\sigma^2(bt)$) ve madde-madde takımı ($\sigma^2(m:t)$) varyans bileşenleri için fark 0,00001 ile 0,00002 arasında değerler almıştır. İki yaklaşım arasındaki varyans bileşenleri değerleri farkının en fazla on binde iki seviyesinde olması aralarında fark olmadığı şeklinde yorumlanabilir. Diğer yandan birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans bileşeni değeri her iki evrenin her koşulu altında iki yaklaşım için en fazla farkın elde edildiği varyans bileşeni olmuştur. Ancak bu durum Genellenebilirlik Kuramı ile Madde Tepki Modellemesinde Genellenebilirlik yaklaşımı arasındaki temel farktan kaynaklanmaktadır. Bu fark; Madde Tepki Modellemesinde Genellenebilirlik yaklaşımının hata varyansını etkileşim varyansından ayırmasıdır. Bir diğer deyişle Genellenebilirlik Kuramı ile kestirilen birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans değeri içinde hata varyansını da barındırır. Bu nedenle MTMG ile elde edilen birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans bileşeni ile hata $\sigma^2(e)$ değerleri toplanıp GK'dan elde edilen birey-madde-madde takımı etkileşimi ($\sigma^2(bm:t)$) varyans bileşeni ile karşılaştırıldığında aradaki farkın 0,00001 ile 0,00002 değerleri arasında olduğu ve bu durumun varyans değerleri arasında fark olmadığı şeklinde yorumlanabileceği görülmüştür. Bu bulgu Briggs ve Wilson'ın (2007) yapmış oldukları MTMG çalışması ile örtüşmektedir.

Genellenebilirlik Kuramı veri seti A ve B evrenleri koşulları altında incelenen bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında iki yaklaşım ile kestirilen değerler açısından fark 0,00001 ile 0,00002 arasındadır. Oldukça küçük olan bu değer madde tepki modelinde genellenebilirlik modeli ve Genellenebilirlik Kuramına göre bağıl ve mutlak hata varyansı, genellenebilirlik ve Phi katsayısı arasında fark olmadığı şeklinde yorumlanabilir. Elde edilen bu sonuç MTMG yaklaşımını ortaya atan çalışmalar ile uyum göstermektedir (Briggs ve Wilson, 2004,2007).

Ayrıca A ve B evrenlerine ait koşullar kendi içlerinde madde takımı ve madde takımlarında yer alan madde sayıları açısından farklılık göstermektedir. Madde takımı sayısı 6 olan birinci ve ikinci koşul arasında madde takımında daha fazla madde bulunan ikinci koşulda daha yüksek güvenirlik elde edilmiştir. Benzer şekilde madde takımı sayısı 9 olan üçüncü ve dördüncü koşullarda madde sayısı fazla olan dördüncü koşulda daha yüksek güvenirlik elde edilmiştir. Madde takımlarında bulunan madde sayıları eşit olan birinci-üçüncü ve ikinci-dördüncü koşullarda madde takımı sayısı fazla olan koşullarda daha yüksek güvenirlik elde edilmiştir. Tüm bu bulgular her dört koşul incelendiğinde toplam madde sayısı arttıkça güvenirliğin arttığını göstermektedir. Araştırmanın bu bulgusu madde takımlarının güvenirliği için yapılan çalışmaları (Thissen, Steinberg ve Mooney, 1989; Sireci, Thissen ve Wainer, 1991; Yen, 1993; Wainer, 1995; Wainer ve Thissen, 1996; Ferrara, Huynh ve Bagli, 1997;

Ferrara, Huynh ve Michaels, 1999; Bradlow, Wainer ve Wang, 1999; Wang, Bradlow ve Wainer, 2002) destekler özelliktedir.

Bunun yanında A ve B evrenleri birey-madde takımı etkileşimi varyansı açısından farklılık göstermektedir. Birey-madde takımı değişkenlik kaynağı (σ^2(bt)) madde takımlarının kolaylık ve zorluk seviyelerinin bireylere göre farklılık gösterip göstermediğinin incelendiği varyans değerlerine sahiptir. Madde takımlarının zorluk seviyelerindeki tutarsızlıkların bireylere göre değişmesi hatayı artırır. Bu nedenle İki evren güvenirlik değerleri açısından karşılaştırıldığında her koşul altında birey-madde takımı etkileşiminin küçük olduğu A evreninde daha yüksek güvenirlik elde edilmiştir. Elde edilen bu sonuç birey madde etkileşimi üzerine yapılmış çalışmalardan elde edilen sonuçlarla örtüşmektir (Alkahtani, 2012; Güler, Kaya Uyanık, Taşdelen Teker, 2012; Hendrickson, 2001; Lee ve Frisbie, 1999; Lee ve Park ,2012; Zhang ve Roberts, 2013).

Araştırmanın temel amacı Genellenebilirlik Kuramına alternatif olarak gösterilen MTMG yaklaşımının çok yüzeyli desenlerde de kullanılabileceğini göstermektir. Elde edilen bulgular sonucunda iki yüzeyli bx(m:t) deseni için MTMG ve GK yaklaşımları arasında varyans değerlerini, mutlak ve bağıl hata varyanslarını ve güvenirlik katsayılarını kestirmede fark olmadığı ortaya çıkmıştır. Bu durumda daha pratik ve kolay analiz yapılan programları olan Genellenebilirlik Kuramının yerine kestirimlerinin daha zor yapıldığı MTMG yaklaşımını kullanmak önerilmemektedir. Ancak MTMG'nin GK karşısındaki en büyük avantajı hata varyansını etkileşim varyansından ayrı kestirebilmesidir. Bu sonuçlar doğrultusunda hata varyansını etkileşim varyansından ayrı olarak kestirebildiği için bx(m:t) deseni için MTMG yaklaşımının kullanılması önerilmektedir.

Çalışmada kullanılan çok yüzeyli desen, maddelerin takımlara yuvalandığı ve bireylerin bunlarla çaprazlandığı rastgele dengelenmiş bx(m:t) yuvalanmış desendir. Çalışmada kullanılan tüm evrenler ve evrenlere ait koşulların tümünde elde edilen güvenirlik değerleri madde sayısı arttıkça güvenirliğin arttığını göstermiştir. Bu nedenle madde takımı ile yapılan çalışmalarda madde takımları ve madde takımlarında yer alan madde sayılarının mümkün olduğunca fazla olması önerilmektedir.

Yapılan çalışma ile daha yüksek güvenirlik elde edilmesi için birey-madde takımı etkileşiminin küçük olması gerektiği sonucuna varılmıştır. Bezer şekilde birey-madde takımı etkisi arttıkça güvenirliğin düştüğü görülmüştür. Elde edilen sonuçlara dayanarak madde takımlarının yer aldığı durumlarda güvenirliği arttırmak için birey-madde takımı etkileşiminin küçük olması; madde takımı etkisinin düşük olması ve genel yetenek ve madde takımı arasındaki ilişki ve madde takımları arasındaki ilişkinin düşük olması önerilmektedir.

MTMG yaklaşımı elde edilen ümit verici sonuçların yanında birbirlerinden farklı iki ölçme kuramını bir arada kullanması açısından değerlidir. Ancak MTMG çalışmaları tek yüzeyli desenler ile sınırlı kalmıştır. MTMG yaklaşımının çok yüzeyli durumlarda nasıl işlediğini bilmek gereklidir. Örneğin MTMG çok yüzeyli modellerde çalışmazsa adres gösterilen sorunlarda GK'nın yerine kullanılabileceği söylenemez. Bu nedenle bu çalışmanın temel amacını Briggs ve Wilson'ın (2004, 2007) yaptıkları çalışmayı tek yüzeyli desenden çok yüzeyli desene çıkartmak oluşturmaktadır. Bu temel amaç doğrultusunda elde edilen bilgiler MTMG yaklaşımının bx(m:t) iki yüzeyli deseni içinde uygun bir yaklaşım olduğunu göstermiştir. Ancak MTMG yaklaşımının daha fazla gelişmesi ve uygunluğunun test edilmesi için farklı sayıda yüzeyin bulunduğu farklı desenlerin kullanılması önerilmektedir.

Son yıllarda yapılan testlerde madde takımlarının kullanımını artması madde takımlarının nasıl puanlanacağı, nasıl analiz edileceği ve madde takımlarının güvenirlik üzerinde etkisinin ne olacağı önemini arttıran bir konu haline gelmiştir. Bu çalışma ile kullanılan maddelerin madde takımları içinde yuvalandığı bx(m:t) deseni için güvenirlik değerlerinin farklı koşullar altında nasıl değiştiği gözlenmiştir. Gözlem sonuçlarında madde takımı ve madde sayısı uzunluklarının; birey-madde takımı varyans değerinin; madde takımı etkisinin; madde takımları arasındaki ilişkinin güvenirlik üzerinde etkili olduğu görülmüştür. Gelecekte yapılacak çalışmalar için madde takımları üzerinde etkisinin olabileceği düşünülen diğer etmenlerinde araştırılması önerilmektedir.

KAYNAKÇA

Alkahtani, S. F. (2012). *Oral performace scoring using generalizability theory and many-facet Rasch measurement: A comparison study* (Unpublished Doctoral Dissertation). The Pennsylvania State University.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*, 364-375.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

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

Briggs, D. C., & Wilson, M. (2004, June). *Generalizability theory in item response modeling*. *Presentation at the International Meeting of the Psychometric Society, Pacific Grove, CA.*

Briggs, D. C., & Wilson, M. (2007). Generalizability theory in item response modeling. *Journal of Educational Measurement, 44*(2), 131-155.

Chien, Y. M. (2008). *An investigation of testlet-based item response models with a random facets design in generalizability theory* (Unpublished Doctoral Dissertation). University of Iowa.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145-168.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27(6),* 440- 458.

Dresher, A. R. (2004, April). An empirical investigation of LID using the testlet model: A further look. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Feldt, L. S., & Quails A. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 105-146). New York: American Council on Education and Macmillan.

Ferrara, S., Huynh, F. L., & Bagli, H. (1997). Contextual characateristics of locally dependent open-ended item clusters on a large-scale performance assessment. *Applied Measurement in Education, 12*, 123-144.

Ferrara, S., Huynh, F. L., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement, 36*, 119-140.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271-288.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models* (Unpublished Doctoral Dissertation). Enschede, University of Twente.

Güler, N., Kaya Uyanık, G. ve Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

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

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100.

Karasar, N. (2004). *Bilimsel araştırma yöntemi* (13. Baskı). Ankara: Nobel Yayınları.

Kim, S. C., & Wilson, M. (2008). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, *10*(4), 408-423.

Kolen, M., & Harris, D. (1987, April). *A multivariate test theory model based on item response theory and generalizability theory*. Paper presented at the American Educational Research Association, Washington, DC.

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

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 9-15.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*(3), 237-255.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30(1), 3-21*.

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

Linacre, J. M. (1999). *FACETS (Version 3.17) [Computer software]*. Chicago: MESA Press.

Lord, F. M. (1983). Unbiased estimation of ability parameters, of their variance, and of their parallel forms reliability. *Psychometrika, 48*, 233-245

Patz, R., Junker, B., Johnson, M. S., & Mariano, L. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics, 27,* 341-384.

Raju, N. S., & Oshima, T. C. (2005). Two prophecy formulas for assessing the reliability of item response theory-based ability estimates. *Educational and Psychological Measurement, 65(3),* 361-375.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika, 53*(3), 349-359.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18,* 229-244.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. USA: SAGE Publications.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26,* 247- 260.

Verhelst N. D., & Verstralen, H. H. F. M. (2001). IRT models for multiple raters. In A. Boomsma, T. Snijders, and M. Van Duijn (Eds.), *Essays in item response modeling* (pp. 89-108). New York: Springer-Verlag.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1-14.

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

Wainer, H., & Wang, C. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37,* 203-220.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). *Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing*. Dordrecht: Kluwer Academic Publishers.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A General bayesian model for testlets: Theory and application. *Applied Psychological Measurement, 26*(1), 109-128.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26,* 283-306.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Zhang, X., & Roberts, W. L. (2013). Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using many-facet Rasch measurement and generalizability theory. *Advances in Health Sciences Education, 18*(5), 929-944.

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56,* 589-600.

EXTENDED ABSTRACT

Introduction

There are three basic theories in education studies which are Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT). There are claims that these theories are not completely different, they can be used together or complementary to each other. Therefore, the researchers are focused on the studies combining the CTT and GT with the IRT. In general, designs in which each item is rated by multiple raters were studied in the body of literature in which GT and IRT were used together. However, these studies were criticized for the fact that the IRT violated the local independence assumption. In order to cope with this issue, other models were worked on in IRT. The integration of the IRT and the GT was first achieved by the studies of Briggs and Wilson (2004, 2007). In the result of these studies, a model named Generalizability in Item Response Modeling (GIRM) has been introduced.

In addition to the promising results obtained, the GIRM approach is also valuable in terms of using the two different measurement theories together, which is carried out in the body of literature very

rarely. However, GIRM studies have remained limited to single facet designs. It is necessary to know how the GIRM approach works in many facet cases. For this reason, the main purpose of this study is to develop the study of Briggs and Wilson (2004, 2007) from single facet design to many facet design.

The many facet design used in the study is a random balanced sx (i: t) nested design in which the items are nested in the testlets and the students are crossed with them. Due to the advantages of the tests composed of testlets the usage of them is increasing in both national and international exams. However, it is evident that when the violations of local item independency of the testlets are ignored, they may cause mistakes in the estimation. In this study, the parameters of the testlets obtained under different conditions were examined.

This study is important with regards to combining different theories. At the same time, the different situations in which the testlets in the study were present were examined within the framework of different theories. For this reason, it is considered important, as it suggests a different approach to parameter estimations in large-scale national tests where the testlets are frequently used.

Method

In this study, the results that were obtained using the Generalizability in Item Response Modeling (GIRM) for different conditions were compared with the results that were obtained using the Generalizability Theory (GT) for the same conditions. The research is a simulation study with regard to deriving the appropriate data by creating controlled conditions. The GT linear dataset sx (i: t) was generated with the balanced random design. Every item was nested in only one testlet. ni and nt varied with different working conditions, but the number of students (ns) was fixed at 1000 for this study. The different datasets were generated for each level of the response of the testlet, testlet length, and the number of testlets. Testlet lengths were determined as 6 and 9; the numbers of the testlets were determined as 3 and 5. The numbers of items used in the study were respectively 18, 30, 27, 45 according to the testlet lengths and the number of the testlets determined. The datasets of the study were produced by R software. Generalizability in Item Response Modeling analyses of the data produced were performed with the WinBUGS software, and Generalizability theory was performed with EDUG software and these estimations were compared for each case.

Results and Discussion

The Generalizability Theory dataset in the study consists of two different populations (A and B), and each population consists of four different conditions. According to Generalizability Theory (GT) and Generalizability in Item Response Modeling (GIRM) model, the difference for estimated student for each condition of A and B populations($\sigma 2(s)$), testlet($\sigma 2(t)$), student testlet interaction($\sigma 2(st)$) and item-testlet ($\sigma 2(i:t)$) variance components have the values between 0,00001 and 0,00002. The difference of variance components' values can be interpreted as there is no difference between the two values.

The difference is between 0,00001 and 0,00002 in terms of the values estimated with two approaches between generalizability and Phi coefficient, relative and absolute error variances analysed under the conditions of A and B populations of GT dataset. This value, which is quite low, can be interpreted as there is no difference between relative and absolute error variance, generalizability and Phi coefficient according to the GT and the GIRM. Moreover, the conditions of the A and B populations differ within themselves in terms of the testlets and the number of items in the testlets. Between the first condition and the second condition with 6 testlets, higher reliability is obtained in the second condition, in which there are more items in the testlet. Similarly, among the third condition and the fourth condition, where the number of the testlet is 9, higher reliability is obtained in the fourth condition in which the number of the items is higher. In the first-third conditions and the second-fourth conditions, where the number of items is equal, higher reliability is obtained in the conditions where the number of items in the testlets is higher. All these findings show that when all four conditions are analysed, the reliability

increases as the total number of items increases. The A and B populations, however, differ in terms of the variance of the student-testlet interaction. The student-testlet variance source (σ2 (st)) has the variance value that is studied whether the simplicity and difficulty levels of the testlets differ according to the individuals. If the inconsistencies in the difficulty of the testlets vary from person to person, the error increases. For this reason, when the two populations are compared in terms of their reliability values, higher reliability is obtained in the population where the individual-testlet interaction is low under all conditions. This result is consistent with the results obtained from studies on individual item interaction.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);33-48

The Scale of Being Able to Say "No" For Children: Validity and Reliability Analysis *

Çocuklar İçin "Hayır" Diyebilme Becerisi Ölçeği: Geçerlik ve Güvenirlik Çalışması

Ferat YILMAZ ** M. Akif SÖZER ***

Öz

Bu araştırmada, çocukların "hayır" diyebilme becerilerini değerlendirmeye yönelik bir ölçeğin geliştirilmesi amaçlanmaktadır. Bu çalışmanın katılımcılarını, HÖ'nün kapsam geçerliğini sağlamak üzere HÖ'de yer alan maddeler hakkında görüşü alınan uzmanlar ile HÖ'nün yapı geçerliği, ölçüm güvenirliği ve madde ayırt ediciliği hakkında fikir edinmek için HÖ'de yer alan maddeleri puanlaması istenen ilkokul 4. sınıf öğrencileri oluşturmaktadır. Bu araştırmada veri toplama aracı olarak sırasıyla HÖ'ye ait 24 maddelik 3'lü Likert, 22 maddelik 3'lü Likert, 12 maddelik 3'lü Likert ve 12 maddelik 5'li Likert olmak üzere dört ayrı taslak form kullanılmıştır. Bu formların geliştirilme süreci, aşağıda ayrıntılı bir biçimde açıklanmaya çalışılmıştır. HÖ'nün kapsam geçerliğini değerlendirmek amacıyla uzman görüşü alınmıştır. Yapı geçerliği için açımlayıcı ve doğrulayıcı faktör analizleri gerçekleştirilmiştir. HÖ'den elde edilen ölçümlerin güvenirliği, Cronbach Alfa ve test yarılama (eşdeğer yarılar) yöntemlerinin kullanılmasıyla analiz edilmiştir. Düzeltilmiş madde toplam korelasyonları ile bağımsız örneklemler için t-testi sonuçlarından yola çıkılarak ölçekte yer alan maddelerin ayırt edicilikleri hakkında bir fikir edinilmiştir. Bu çalışmanın sonuçları, çocukların "hayır" diyebilme becerileri ile ilgili yapılacak çalışmalarda HÖ'nün geçerli ve güvenilir sonuçlar verebilecek bir ölçme aracı olarak kullanılabileceğini ortaya koymuştur.

Anahtar Kelimeler: "Hayır" diyebilme, "hayır" diyebilme ölçeği, ölçek geliştirme

Abstract

In this study, it is aimed to develop a scale for assessing children's ability to say "no". The participants included in the study are the area experts that helped to achieve content validity, and fourth graders who are asked to score the items in the SN to have an idea about the construct validity, measurement reliability and item discrimination of the item in the scale. In this study, four different drafts of the SN which contained 24 items of 3-pointed Likert type, 22 items of 3-pointed Likert type, 12 items of 3-pointed Likert type and 12 items of 5-pointed Likert type were used. In order to evaluate the content validity of the SN, the experts' opinions were received. For the construct validity, exploratory and confirmatory factor analyses were conducted. Reliability of measurements obtained with the SN was analyzed by using the methods of Cronbach's Alpha and Split-half. An opinion was formed about the distinctiveness of items in the scale setting out from the corrected item total correlations (CITC) and the results of the independent samples t-tests. The results of the study showed that the SN can be used as a data collection tool, and is able to produce valid and reliable interpretations in studies that aim to explore children's ability to say "no".

Keywords: Being able to say "no", the scale of being able to say "no", scale development

* This study is derived from a part of Ph.D. dissertation by Ferat YILMAZ entitled The Investigation of Primary School 4th Grade Students' Skills of Saying "No", submitted to Gazi University under the supervision of Assoc. Prof. Dr. M. Akif SÖZER.

** Res. Asist. Dr., Dicle University, Ziya Gökalp Faculty of Education, Diyarbakır-Turkey, ferat.yilmaz@dicle.edu.tr, ORCID ID: http://orcid.org/0000-0002-4947-5416

*** Assoc. Prof. Dr., Gazi University, Gazi Faculty of Education, Ankara-Turkey, akif@gazi.edu.tr, ORCID ID: https://orcid.org/0000-0002-1291-4067

DOI: 10.21031/epod.350631
Geliş Tarihi:11.10.2017
Kabul Tarihi:10.02.2018

INTRODUCTION

The world is "biologically livable" for all living organisms unless its mechanism for natural balance is disturbed. The concept of livability needs assessment beyond its biological sense when a human is concerned; because livability has several aspects such as emotional, mental, social, cultural, economic and political aspects apart from biological. One of the most important instruments for making the world a livable place for everybody is people's rights. Every individual has basic rights which make the world livable for them due to the fact that they are human beings. Of those rights, the ones come to mind initially are personal, political, social and economic rights. However, individuals' rights are not limited to these.

Each individual has rights stemming from the situation they are in, apart from their personal, political, social and economic rights. Those rights- which are not written or guaranteed by the law- are called "assertiveness rights" (Garner, 2012). Assertiveness rights are the rights changing our behaviors and ideas, allowing to take on responsibility only in matters we wish to choose, allowing to make mistakes and saying 'I don't know', 'I don't understand' or 'it is none of my business (Smith, 1998). Rights such as being treated equally and respectably, being committed to various values, privacy, observing individual needs and deciding how to behave in a given situation are also among those rights (Bishop, 2010). Finally, determining personal priorities, questioning injustice, refusal recommendations, refusal to be confirmed by others and being able to say "no" without feeling guilty are also considered being assertiveness rights (Pfeiffer, 2010).

Even though being able to say "no" is an assertiveness right, it is not usually displayed by every individual. This situation is thought to be related with the fact that being able to say "no" is an ability. This ability is defined as the process of creating the capacity to refuse risky behaviors with one's own will and with his/her choice by Aslan and Özcebe (2008). This study, however, considers this ability as one's ability to refuse demands, offers and behaviors targeting him/her by saying "no", and to resist against potential manipulation efforts in cases where his/her personal rights and limits are violated or open to be violated. Considering the ability to say "no" in this way stems from necessity to implement the decision at the stages of refusal and resistance after a decision has been made to put the ability into action.

Refusal indicates a natural response to something undesired to be done at the relevant time (Bragger, 1982). An individual is asked to state that he/she does not accept the demand by using the most appropriate verbal expressions and body language consistent with those expressions and to say "no" at the stage of refusal. Yet, an individual's behavior of refusal can cause manipulation efforts such as insistence from the opposite side, as addressed by Masterson (2011). Therefore, an individual is expected not to leave his/her decision to change for manipulation effort probable to come from the opposite side at the stage of refusal.

"No" should be said politely but clearly when refusing (Blair, 2008). Shortly speaking, sincere explanation should be given in relation to "no" without giving an excuse (Bolton, 1979). Stating a long explanation and/or giving an excuse may lead the other person to conclude that the stated reasons are insufficient or to create an opportunity for his/her demands (Breitmen and Hatch, 2011). Besides, rationalized and complicated explanations may make both sides feel bad, and result in repeating the demand. Therefore, saying "no" should include only an unconditional behavior of saying "no", expressing the feelings honestly and briefly (Holland and Ward, 1990). No permission should be asked to say "no" because asking for permission may mean that somebody else, not the individual himself/herself, is responsible for individual's behavior. The individuals to whom the responsibility is transferred can also hinder displaying the behavior of saying "no" (Paterson, 2000). For this reason, the responsibility of saying "no" should be taken on by the individual, and not be left to the opposite side. Also, the others on the other side should not be permitted to judge the decision when one says "no" (Clark, 2003) since those judgments might cause pressure on the individual to accept the demand. The phrase "may be" or other phases meaning "may be" should not be used when one wants to say "no". The phrase "may be" can mostly be understood as a potential "yes". This might lead the opposite side to increase the hope. Being refused after a "maybe" answer might also cause bigger disappointment on the refused person (Beagrie, 2007). Assertive body language should be used when

saying "no" at the stage of refusal. A decisive, but with a polite tone of voice, answer should be used for this. The voice should not tremble, and should not be in an apologizing manner. The truth should be stated in a neutral rather than the defensive tone of voice. The body should be in an upright posture and eye contact should be set up (Deering, 1996). Nonverbal messages should be conveyed accurately. If "no" is said, a smiling facial expression to lead the opposite side to insist on his/her demand should be avoided (Moon, 2009).

One should be decisive in front of the insistence of the person making demands at the stage of resistance. Too much time should not be spent in that place, and movement should be made to another place, and if these are not possible, the topic of speaking should be changed in order not to be persuaded by the person who is refused by saying "no" (DeJong, 1986). If that does not work, one should go on with what he is doing when there is a demand (Rees and Graham, 1991). The previously said "no" should be repeated without making any new explanations whenever there is a demand (Deering, 1996). If others' demands have been accepted for years, it should not be expected that they will accept the behavior of refusal at the first time. More efforts should be made in this respect and the opposite side should be refused more strongly (Paterson, 2000). The questions about why "no" is said should not be answered. Insult and implications should not be responded in order not to distract attention (Clark, 2003). Assertiveness rights should be used against such manipulating behaviors as praise, hypocrisy, deceiving by saying nice words, putting someone in a difficult situation, despising, accusing, threatening, casting aspersion on someone and emotional exploitation (Dalley, 2013; Potter, 2007; Potts and Potts, 2013).

The ability to say "no", which is put into action at the stages of refusal and resistance, is important in that it can protect children against using substance (Tokur-Kesgin, 2012), sexual abuse (Elliot, Browne and Kilcoyne, 1995; Lecler, Wortley and Smallbone, 2011), having problems in time management (Mackenzie and Nickerson, 2009), moral violations (Leming, 1997; Szpalski, Gunzburg and DeKleuver, 2003) and against being exposed to online risks (Bal and Kahraman, 2015). For this reason, this ability should be developed at early ages. While grades 1-3 are considered as the most appropriate stages, grades 4 and 6 are thought to be the periods before too late for developing the ability (Hermann and McWhirter, 1997). Taking early steps in this respect can also pave the way for the subsequent developmental periods when such abilities gain more importance (Belgrave, Reed, Plybon and Corneille, 2004; Scheier, Botwin, Diaz and Griffin, 1999). However, on reviewing the literature about whether or not such a way is paved, no studies considering children's ability to say "no" directly and in great details were found. Instead, it was found that the ability was considered in various studies (Durualp and Aral, 2010; Gündoğdu, 2012; Tuna Özcivanoğlu, 2010) within the scope of adults' assertiveness skills in a restricted manner. Probably, one of the reasons for this is that there is no data collection tool to use for conducting research. Thus, this study aims to develop a scale for assessing children's ability to say "no".

METHOD

This is a study for developing a scale. The participants, data collection tools, the process of developing the scale of being able to say "no" (SN), and the data analysis were described below.

Participants

The participants included in this study were the experts whose opinions were consulted for the items in the SN to attain content validity and fourth graders who are asked to score the items in the SN to have an idea about the construct validity, measurement reliability and item discriminations of the scale. First, the expert opinion was consulted for the draft forms of the SN for qualitative and quantitative evaluations. The views of six experts, the three were the instructors in the primary education department of Gazi Faculty of Education of Gazi University and the three were the instructors in the departments of educational management, guidance, and psychological counseling, and curriculum and instruction, were obtained for the first draft form in terms of quality. Having made the necessary

corrections, the second draft was shown eight experts for its quantitative evaluation. The seven out of eight was from Dicle University Ziya Gökalp Educational Faculty (that is to say, two instructors in primary education department, one instructor in curriculum and instruction department, and one instructor in each of measurement and evaluation, social studies teaching and mathematics and science teaching departments) and the last one was from Inönü University Educational Faculty (instructor in primary education department).

Pre-interviews were conducted with ten students who were fourth graders by using the drafted form prepared on the basis of expert opinion. Two pre-applications were conducted for Exploratory Factor Analysis (EFA) to test construct validity following the pre-interviews. According to Güngör (2016), researchers who want to test a CFA-specified construct with EFA should test the same model in a different sample to find the best model. This is the ideal situation recommended by some researchers (Kılıç-Çakmak, Çebi, and Kan, 2014). Thus, one re-application was conducted for Confirmatory Factor Analysis (CFA). The first application was done with 302 fourth graders (who were in the 9-11 age range) in a primary school which could be considered as socioeconomically at the medium level. The second application was done with 250 fourth graders (in the 9-11 age range) in a different primary school of similar properties. The data necessary for CFA were also collected from 230 students of fourth graders (9-11 age range) attending a primary school at a medium socioeconomic level. The socioeconomic levels of the schools in which these applications were carried out have been defined based on the descriptions of school administrators, as the provincial directorate of education does not classify the socioeconomic levels of schools.

And finally, the main applications were performed with 907 students of fourth grade attending 13 different primary schools in the central districts of Diyarbakır which were selected through random sampling. Efforts were made to interpret the measurement reliability of the SN in terms of refusal and resistance and the discriminating levels of the items in the scale basing on the findings obtained from these applications.

Data Collection Tool

In this study, four different drafts of the SN which contained 24 items of 3-pointed Likert type, 22 items of 3-pointed Likert type, 12 items of 3-pointed Likert type and 12 items of 5-pointed Likert type were used respectively. The development process for these forms is described in details below.

Development Process and Data Analysis

The first stage in the development process of the SN was defining the scale structurally. Accordingly, the refusal dimension of the scale measured whether or not students could say “no” for the demands and behaviors they did not like or found unreliable. On the dimension of resistance, on the other hand, efforts were made to determine whether or not students stepped back due to their feelings they can have after saying “no” or manipulation efforts they encounter.

Having described the structures in the SN, a 24-item draft of 3-pointed Likert type was created to measure the ability to say “no” on the dimensions of refusal and resistance. The draft was presented to six experts teaching at the Gazi University, Gazi Faculty of Education for assessment in terms of quality. After their views, five items were removed from the scale since they were found to be inappropriate, and the remaining items were modified to improve their clarity and finally, three items were added to the scale based on the experts’ suggestions. After making those corrections to the scale, expert opinion was consulted again according to Davis (1992) technique to attain the content validity of the 22-item draft of 3-pointed Likert type. Based on the technique, seven experts from Dicle University Ziya Gökalp Educational Faculty and one expert from Inönü University were asked to assess the 22 items planned to be included in the scale as “4: highly relevant, 3: quite relevant, 2: somewhat relevant, and 1: not relevant”. Following the evaluation, the decision was made to keep all of the 22 items in the scale as they were, and a pre-interview was conducted with 10 students in fourth

grade. Since it was found that the students were able to understand all the statements in the 22 items included in the draft form, the application necessary for EFA was started without making any changes.

The first pre-application for EFA was conducted with 302 fourth graders going to a primary school which might be considered to be socioeconomically at the medium level. In consequence, the Kaiser-Meyer-Olkin (KMO) value was found as 0.80, and the Barlett test result was found as significant ($p<0.05$). The EFA results indicated that all of the 22 items, whose content validity was determined on the basis of expert opinion, could be grouped in the factors of refusal and resistance as it had been predicted. The SN explained 31% of the total variance it that form. Although the ratio is considered adequate for social sciences (Bayram, 2015), it is necessary to go well beyond this proportion in multi-factor structures (Büyüköztürk, 2012). For this reason, the items whose common factor variance was smaller than 0.3 (Pallant, 2007) were excluded, and the EFA was re-done with the remaining 12 items. In consequence, it was found that the KMO value was 0.76 and the Barlett test result was significant ($p<0.05$). In this form, the SN explained 41.28% of the total variance. Since this proportion is bigger than the unexplained variance ratio (58.72) (Seçer, 2013), it was unacceptable. We believed that this was due to the fact that the scale was 3-pointed Likert type, therefore, the scale was restructured by using 5-pointed Likert type, and then tested again for construct validity with 250 students. The orthogonal rotation technique was used during EFA, because no correlation (Pallant, 2007) between refusal and resistance dimensions was hypothesized. Accordingly, the KMO value was found as 0.84 and the Barlett test was found as significant ($p<0.05$). These results showed that the data obtained from the application for EFA fitted factor analysis, and they also indicated that the number of participants for EFA was adequate (Seçer, 2013). Additionally, the skewness values for the refusal and resistance dimensions were -.21 and 27, respectively. This means that the scores obtained within the dimensions of refusal and resistance sub-scales were normally distributed (Büyüköztürk, 2012). Thus, we met the assumptions the EFA requires. We also investigated whether or not the factor structure obtained from EFA was confirmed by first level CFA. The data collected from 230 fourth graders demonstrated that the CFA model confirmed the 12-item and 2-factor structure of the SN obtained from EFA.

After pre-applications concerning construct validity, the main application was done, and the reliability of measurements conducted with the SN was analyzed by using the methods of Cronbach's Alpha and Split-half. Finally, item analysis was done with the data come from the main application in the process of SN development, and an opinion was formed about the distinctiveness of items setting out from the corrected item total correlations (CITC) and the results of the independent samples t-tests which had been administered to determine whether there were any significant differences between the each items' scores of the top and the bottom 27% groups determined according to participants sub-scale score averages.

FINDINGS

Findings concerning the content and construct validity of the SN, reliability of measurements and item analysis as well as the interpretations of the findings are presented below.

Content Validity

Table 1 below shows the validity indices calculated on the basis of expert opinion about whether or not the items in the 22-item draft form of SN are related to the latent variable of being able to say "no" according to Davis (1992) technique.

Table 1. SN Content Validity Indices

Item	1th expert	2nd expert	3rd expert	4th expert	5th expert	6th expert	7th expert	8th expert	Validity Indices (%)	Item	1th expert	2nd expert	3rd expert	4th expert	5th expert	6th expert	7th expert	8th expert	Validity Indices (%)
I1	4	4	4	4	3	4	4	4	100	I12	4	4	4	4	3	4	4	3	100
I2	4	4	4	4	4	4	4	4	100	I13	3	4	4	4	4	4	4	3	100
I3	4	4	4	3	4	4	4	4	100	I14	4	4	4	3	4	4	4	4	100
I4	3	4	3	4	4	3	3	4	100	I15	3	4	4	4	4	4	4	4	100
I5	4	4	4	4	4	4	4	4	100	I16	3	3	3	3	4	4	4	4	100
I6	3	4	4	4	3	4	4	3	100	I17	3	3	4	4	4	4	4	4	100
I7	4	4	4	4	4	4	4	4	100	I18	4	1	4	4	4	3	4	4	88
I8	4	4	4	4	4	4	4	4	100	I19	4	4	3	3	4	4	4	3	100
I9	4	4	4	4	4	4	4	4	100	I20	4	4	4	3	4	4	4	3	100
I10	3	4	4	4	4	4	4	4	100	I21	3	4	4	4	4	4	4	4	100
I11	4	4	4	2	4	4	4	4	88	I22	3	3	4	4	4	4	4	3	100

4: highly relevant, 3: quite relevant, 2: somewhat relevant

As is clear from Table 1, the validity indices calculated for all items in the SN through the formula [number(4) + number(3)/total number of experts] range between 88% and 100%. Since the indices were above 80%, it was found that all the items in the scale were directly related to property to be measured- that is to say, they had a high content validity (Davis, 1992).

Construct Validity

Findings concerning the exploratory and confirmatory factor analysis performed to test the construct validity of the SN are described below.

Exploratory Factor Analysis

Table 2 shows the factor loadings of the items in both dimensions, the common factor variance (CFV) for each item, the percentage of variance that the two dimensions explain (PEV) on their own and the percentage of variance they explain in total.

Table 2. EFA Factor Loadings of the items in SN and Common Factor Variances

Item	Refusal	Resistance	CFV	PEV
Item 1	**.65**	-.12	.39	
Item 2	**.73**	-.08	.42	
Item 4	**.71**	-.04	.42	26.99%
Item 7	**.65**	-.09	.39	
Item 10	**.73**	-.03	.43	
Item 11	**.71**	-.02	.33	
Item 14	-.01	**.69**	.34	
Item 16	-.04	**.75**	.36	
Item 19	-.09	**.80**	.52	24.64%
Item 20	-.06	**.74**	.55	
Item 21	-.07	**.65**	.37	
Item 22	-.15	**.75**	.45	
Total PEV				51.63%

As is clear from the EFA results shown in Table 2, the factor loadings for items available on the dimension of refusal range from 0.65 to 0.73 and the factor loadings of items on the dimension of resistance range from 0.65 to 0.80. While all of these items have the stated factor loadings on their own dimension, they have factor loadings below 0.30 on the other dimension. This situation indicates

that the items are related to the dimensions on which they are predicted to belong and that they do not overlap. The reason for this is that there are differences bigger than 0.10 (Akbulut, 2010) between the factor loadings of the items in the SN on their own dimension and the factor loadings on the other dimension. Besides, factor loadings of the items in the SN ranging from 0.65 to 0.80 mean a high proportion for social sciences (Büyüköztürk, 2012). The SN is capable of explaining 51.63% of total variance with its structure of 5-pointed Likert type. As expected, it is observed that the proportion is higher than the proportion of unexplained variance (48.37%).

Confirmatory Factor Analysis

Table 3 shows the t values for items in the structure obtained through first level CFA.

Table 3. The t Values for Items in the SN

Refusal	*t*	Resistance	*t*
I1	9.41	I7	6.93
I2	9.66	I8	11.27
I3	7.05	I9	10.67
I4	8.88	I10	12.18
I5	7.01	I11	5.61
I6	7.08	I12	11.32

The t values shown in the Table 3 range from 7.01 to 9.66 for the dimension of refusal whereas, range from 5.61 to 12.18 for the dimension of resistance. Since the values were above 2.56, they were significant at the level of 0.01. This indicated that the latent variables in the scale predict all the items in the scale. Although these results show a desirable situation for the construct validity of SN, the goodness of fit indices should also be analyzed to assess for the structure to be acceptable as a whole (Şimşek, 2007). Table 4 shows some of the goodness of fit indices analyzed in this context.

Table 4. CFA Fit Indices for SN

	First Order CFA	Evaluation
p	**0.001**<0.05	Significant
x^2 /df	90.1/53=**1.7**≤2	Perfect fit
RMSEA	0.05≤**0.057**≤0.08	Acceptable fit
CFI	0.95≤**0.95**≤1	Perfect fit
NFI	0.90≤0.**91**≤0.95	Acceptable fit
NNFI	0.90≤0.**94**≤0.95	Acceptable fit
IFI	0.95≤**0.96**≤1	Perfect fit
SRMR	0.05≤**0.054**≤0.08	Acceptable fit
PNFI	0.50≤**0.73**≤0.95	Acceptable fit
PGFI	0.50≤**0.64**≤0.95	Acceptable fit
GFI	0.90≤**0.94**≤0.95	Acceptable fit
AGFI	0.90≤**0.90**≤1	Perfect fit

The first index that should be focused on in Table 4 is Chi square (badness of fit index, x^2) statistics-which is known as goodness of fit statistics. The value obtained in consequence of chi square should be significant at the level of 0.05. The chi square statistics in this study is significant at the level of 0.05. Since this means that there is a significant difference between observed data matrix and expected data matrix, it is necessary to analyze other fit indices. Of the other fit indices which were analyzed from the aspect of first level CFA x^2/df (chi square/degree of freedom), CFI (Comparative Fit Index), IFI (Incremental Fit Index) and AGFI (Adjusted Goodness of Fit Index) represented perfect fit. NFI (Normed Fit Index), NNFI (Non-normed Fit Index), SRMR (Standardized Root Mean Square Residual), PNFI (Parsimonious Normed Fit Index) and GFI (goodness of fit index), on the other hand, indicated good or acceptable fit (Bayram, 2013, Çokluk, Şekercioğlu and Büyüköztürk, 2012; İlhan

and Çetin, 2013; Seçer, 2013). Factor loadings found in structures obtained through CFA are shown in Figure 1 below.

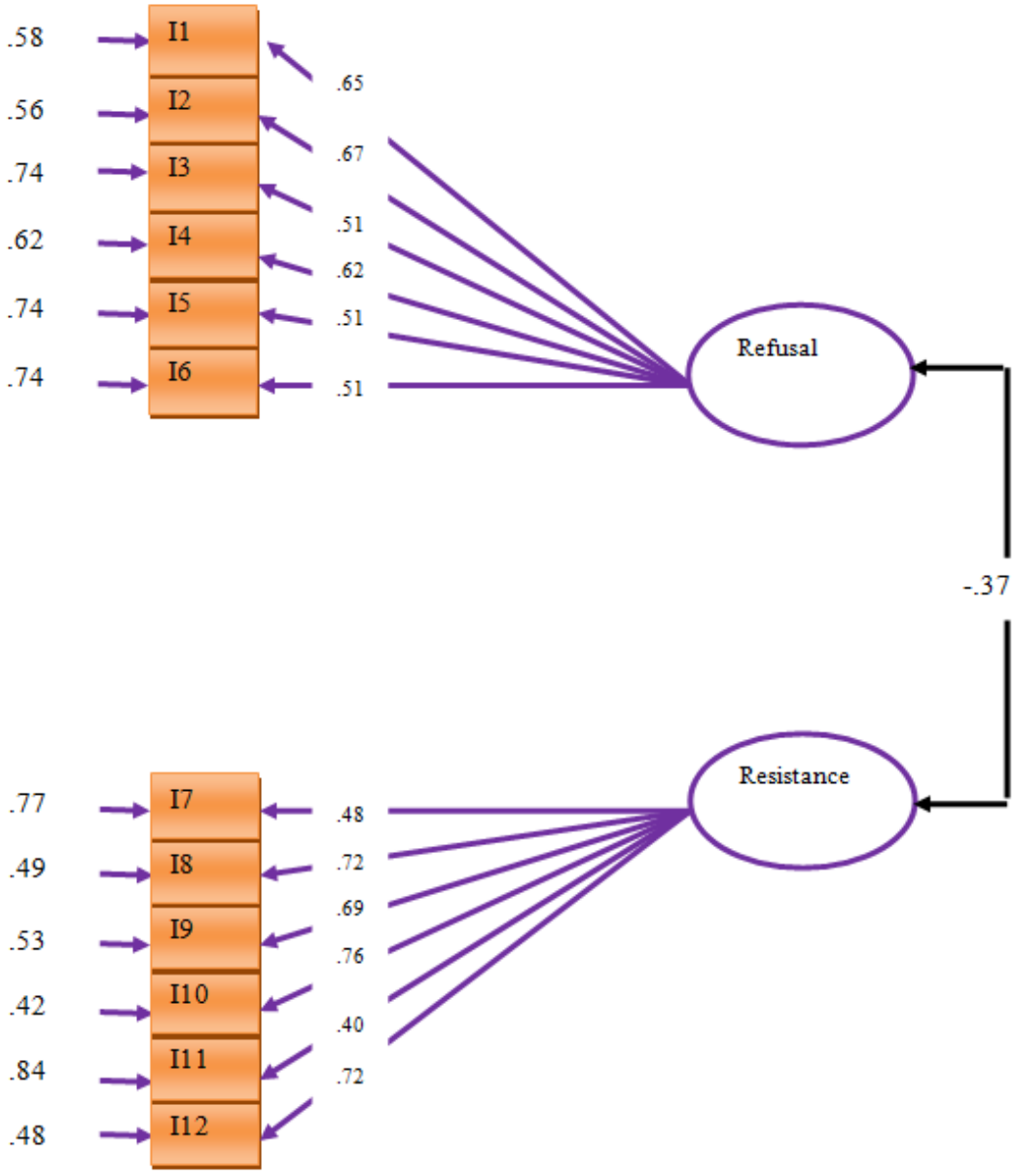


Figure 1. (CFA) Factor Loadings for SN

As it is clear from Figure 1, the factor loadings of SN on the dimension of refusal range between 0.51 and 0.67. The factor loadings for items on the dimension of resistance range from 0.40 to 0.76. Because all of these values were above 0.30, all of the 12 items in the SN had sufficient factor loadings.

Measurement Reliability

The Cronbach's Alpha and Split-half methods were conducted so as to make comments on the reliability of measurements performed with SN in this study. The findings obtained in this respect are shown in Table 5.

Table 5. Measurement Reliability for SN

	Cronbach's Alpha	Split-half
Refusal	0.78	0.72
Resistance	0.77	0.77

As evident from Table 5, the coefficients found by Cronbach's Alpha method for measurements performed with SN were 0.78 and 0.72 for the dimensions of refusal and resistance, respectively. On the other hand, the Spearman-Brown coefficient found by Split-half method is 0.72 for the dimension of refusal and it is 0.77 for the dimension of resistance. Since all of these coefficients are above 0.70 (Sipahi, Yurtkoru and Çinko, 2010), measurements with SN can be said to be reliable on the dimensions of refusal and resistance.

Item Discrimination

CITC was calculated for items in order to find whether or not the items available on the sub-dimensions of refusal and resistance were discriminating, and Independent Sample *t*-Test was performed by considering item scores of individuals in 27% top and bottom groups. The results obtained are shown in Table 6.

Table 6. Item Discrimination for Items in the SN

			İndependent Sample *t* Test						
		CITC	Group	N	X	sd	*t*	df	p
Refusal	I1	0.51	Bottom	245	2.46	1.22	-28.27	321.02	0.000
			Top	245	4.85	0.49			
	I2	0.54	Bottom	245	2.58	1.51	-24.25	254.90	0.000
			Top	245	4.96	0.23			
	I3	0.54	Bottom	245	2.73	1.64	-21.02	253.21	0.000
			Top	245	4.95	0.23			
	I4	0.49	Bottom	245	2.50	1.32	-23.82	341.42	0.000
			Top	245	4.70	0.60			
	I5	0.55	Bottom	245	2.46	1.29	-28.59	275.03	0.000
			Top	245	4.89	0.33			
	I6	0.50	Bottom	245	2.42	1.39	-26.52	282.15	0.000
			Top	245	4.86	0.39			
Resistance	I7	0.47	Bottom	245	2.84	1.48	-22.28	254.74	0.000
			Top	245	4.97	0.22			
	I8	0.52	Bottom	245	3.04	1.56	-19.46	248.08	0.000
			Top	245	4.99	0.14			
	I9	0.60	Bottom	245	2.36	1.32	-28.69	284.45	0.000
			Top	245	4.88	0.38			
	I10	0.58	Bottom	245	2.24	1.23	-31.51	298.28	0.000
			Top	245	4.85	0.41			
	I11	0.35	Bottom	245	2.66	1.42	-21.40	325.80	0.000
			Top	245	4.76	0.59			
	I12	0.57	Bottom	245	2.64	1.48	-23.92	263.24	0.000
			Top	245	4.93	0.29			

According to Table 6, CITCs for the items on the dimension of refusing range from 0.49 to 0.55 whereas CITCs for the items on the dimension of resistance range from 0.35 to 0.60. The fact that the values are above 0.30 shows that the relevant items are in relation to the dimension to which they belong (Akbulut, 2010) and that they have discriminating power (Büyüköztürk, 2012). Another proof may be that there are significant differences between the scores individuals in the 27% top and bottom groups ($p<0.001$).

RESULTS AND DISCUSSION

SN is a self-report scale intending to measure fourth graders' ability to say "no". The scale is composed of refusal and resistance dimensions. Each dimension contains six items. The students are asked to score each statement in the items as "Never" "Rarely", "Sometimes", "Mostly" and "Always" in 5-pointed Likert type. In assessing the scores received from the SN, coding for the dimension of refusal is rated as it is while the coding for the dimension of resistance is rated reversely.

The findings obtained in this study indicated that the SN was a valid tool of measurement in measuring fourth graders' ability to say "no". The reliability tests employed in this study demonstrated that reliable measurements could be done with SN. Item analyses also suggested that the individuals answering the SN could be discriminated in terms of their performance in saying "no". These results have proven that the SN can be used as a measurement tool capable of yielding valid and reliable results in studies to be conducted in relation to children's ability to say "no".

An examination of other measurement tools, measuring the ability to say "no" (García-Ros, Pérez-González, & Hinojosa, 2004; Macan, Shahani, Dipboye, & Philips, 1990; Scheier et al., 1999) makes it clear that they evaluate the ability to say "no" in scales developed for different purposes, in limited ways and at item level, or they aim to measure the ability only in such contexts as substance use and time management. This current study, however, aims to contribute to the relevant literature with its data collection tool which can measure the ability in details on the dimensions of refusal and resistance and which can be used in several contexts.

However, this study has some restrictions. Within its scope, this study collected the data only from fourth graders who were in the 9-11 age range. Therefore, it is suggested that the reliability and validity of the SN should be conducted with different age groups and individuals of different developmental periods in the future studies. Thus, it would be possible to see how individuals' ability to say "no" varies at different periods of development. While the content and construct validity of SN was analyzed in this study, its concurrent validity was not analyzed. It is believed that conducting concurrent validity analysis would be beneficial. It is recommended in this context that the data collection tools to measure assertiveness skills and social skills which can be related to the ability to say "no" (Caldarella and Merrell, 1997; Eslami, Mazaheri, Mostafavi, Abbasi and Noroozi, 2014; Gresham and Elliot, 1990; Vaz, Parsons, Passmore, Andreou and Falkmer, 2013) should be employed. This current study has determined the measurement reliability of SN by using the Cronbach's Alpha and Split-half methods. We believe that test-retest method could also be used in analyzing the measurement reliability of SN in prospective studies. This would allow us to see whether or not the scale performs consistent measurements at different times. In this study, only the Turkish version of the scale was tested for validity and reliability. The validity and reliability of the English version of this scale can also be studied in future studies.

REFERENCES

Akbulut, Y. (2010). *Sosyal bilimlerde SPSS uygulamaları.* İstanbul: İdeal Kültür Yayıncılık.

Aslan, D. ve Özcebe, H. (2008). *Eğitim kurumlarında sigarasızlık politikaları.* Ankara: Klasmat Matbaacılık.

Bal, P. N., & Kahraman, S. (2015). The effect of cyber bullying sensibility improvement group training program on gifted students. *Journal of Gifted Education Research, 3*(1), 48-57.

Bayram, N. (2013). *Yapısal eşitlik modellemesine giriş: AMOS uygulamaları*. Bursa: Ezgi Kitabevi.

Bayram, N. (2015). *Sosyal bilimlerde SPSS ile veri analizi.* Bursa: Ezgi Kitabevi.

Beagrie, S. (2007). How to say no. *Occupational Health, 59*(8), 3-3.

Belgrave, F. Z., Reed, M. C., Plybon, L. E., & Corneille, M. (2004). The impact of a culturally enhanced drug prevention program on drug and alcohol refusal efficacy among urban African American girls. *Journal of Drug Education, 34*(3), 267-279. doi:10.2190/H40Y-D098-GCFA-EL74

Bishop, S. (2010). *Develop your assertivneness.* London: Kogan Page.

Blair, H. (2008). Politely say no to personal or professional requests. *ONS Connect, 23*(8), 23-23.

Bolton, R. (1979). *How to assert yourself, listen to others, and resolve conflicts.* New York: Simon & Schuster, Inc.

Bragger, J. D. (1982). Responses to teacher commands: An effective teaching strategy. *The Modern Language Journal, 66*(1), 9-12. doi:10.2307/327809

Breitman, P., & Hatch, C. (2011). *How to say no without feeling guilty.* London: Vermilion.
Büyüköztürk, Ş. (2012). *Sosyal bilimler için veri analizi el kitabı.* Ankara: Pegem Akademi.
Caldarella, P., & Merrell, K. (1997). Common dimensions of social skills of children and adolescents: A taxonomy of positive behaviors. *School Psychology Review, 26*, 264-278.
Clark, C. C. (2003). *Holistic assertiveness skills for nurses: Empower yourself and others.* New York: Springer.
Çokluk, Ö., Şekercioğlu, G. ve Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları.* Ankara: Pegem Akademi.
Dalley, D. (2013). *Developing your assertiveness skills.* Lancashire: Universe of Learning.
Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research, 5*(4), 194-197. doi:10.1016/S0897-1897(05)80008-4
Deering, C. G. (1996). Working with people: Learning to say no. *American Journal of Nursing, 96*(4), 62-64. doi:10.2307/3465096
DeJong, W. (1986). *Project DARE: Teaching kids to say "no" to drugs and alcohol.* Washington, DC: National Inst. of Justice.
Durualp, E., & Aral, N. (2010). A study on the effects of play-based social skills training on social skills of six-year-old children. *Hacettepe University Journal of Education*, 39, 160-172.
Elliott, M., Browne, K., & Kilcoyne, J. (1995). Child sexual abuse prevention: What offenders tell us. *Child Abuse & Neglect, 19*(5), 579-594. doi:10.1016/0145-2134(95)00017-3
Eslami, A. A., Mazaheri, M. A., Mostafavi, F., Abbasi, M. H., & Noroozi, E. (2014). Farsi version of social skills rating system-secondary student form: Cultural adaptation, reliability and construct validity. *Iranian Journal of Psychiatry and Behavioral Sciences, 8*(2), 97-104.
García-Ros, R., Pérez-González, F., & Hinojosa, E. (2004). Assessing time management skills as an important aspect of student learning: The construction and evaluation of a time management scale with Spanish high school students. *School Psychology International, 25*(2), 167-183. doi:10.1177/0143034304043684
Garner, E. (2012). *Assertiveness: Re-claim your assertive brightright*, Retrieved from http://www.gesp.ipg.pt/files/assertiveness.pdf
Gresham, F. M., & Elliot, S. N. (1990). *Social skills rating system.* MN: American Guidance.
Gündoğdu, R. (2012). Effect of the creative drama-based assertiveness program on the assertiveness skill of psychological counsellor candidates. *Educational Sciences: Theory & Practice, 12*(2), 677-693.
Güngör, D. (2016). A guide to scale development and adaptation in psychology. *Turkish Psychological Articles, 19*(38), 104-112.
Herrmann, D. S., & McWhirter, J. J. (1997). Refusal and resistance skills for children and adolescents: A selected review. *Journal of Counseling & Development*, 75, 177-187. doi:10.1002/j.1556-6676.1997.tb02331.x
Holland, S., & Ward, C. (1990). *Assertiveness: A practical approach.* Oxon: Speechmark.
İlhan, M., & Çetin, B. (2013). The validity and reliability study of the Turkish version of an online learning readiness scale. *Educational Technology: Theory and Practice, 3*(2), 72-101.
Kersey, K. C., & Masterson, M. L. (2011). Learn to say yes! When you want to say no! To create cooperation instead of resistance. *Young Children, 66*(4), 40-44.
Kılıç Çakmak, E., Çebi, A., & Kan, A. (2014). Developing a "Social Presence Scale" for e-learning environments. *Educational Sciences: Theory & Practice, 14*(2), 755-768.
Lecler, B., Wortley, R., & Smallbone, S. (2011). Victim resistance in child sexual abuse: A look into the efficacy of self-protection strategies based on the offender's experience. *Journal of Interpersonal Violence, 26*(9), 1868-1883. doi:10.1177/ 0886260510372941
Leming, J. S. (1997). Whither goes character education? Objectives, pedagogy, and research in education programs. *Journal of Education, 179*(2), 11-34.
Macan, T. H., Shahani, C., Dipboye, R. L., & Philips, A. P. (1990). College students' time management: Correlations with academic performance and stress. *Journal of Educational PSychology, 82*(4), 760-768. doi:10.1037/0022-0663.82.4.760
Mackenzie, A., & Nickerson, P. (2009). *The time trap.* New York: American Management Association.
Moon, J. (2009). *Achieving success through academic assertiveness: Real life strategies for today's higher education students.* New York: Routledge.
Pallant, J. (2007). *SPSS survival manual: A step by step guide to data analysis using SPSS for windows*. Berkshire: Open University.
Paterson, R. J. (2000). *How to express your ideas and stand up for yourself at work and in relationships.* Oakland: New Harbinger.
Pfeiffer, R. (2010). *Relationships: Assertiveness skills.* Sedona, AZ: Growth.
Potter, J. V. (2007). *Assertiveness, individuation & autonomy.* Redding: Jubilee Enterprises AFS.
Potts, C., & Potts, S. (2013). *Assertiveness: How to be strong in every situation.* Cornwall: TJ International.
Rees, S., & Graham, R. S. (1991). *Assertion training: How to be who you reaaly are.* London: Routledge.

Scheier, L. M., Botvin, G. J., Diaz, T., & Griffin, K. W. (1999). Social skills, competence, and drug refusal efficacy as predictors of adolescent alcohol use. *Journal of Drug Educatıon, 29*(3), 251-578. doi:10.2190/M3CT-WWJM-5JAQ-WP15

Seçer, İ. (2013). *SPSS ve LISREL ile pratik veri analizi.* Ankara: Anı Yayıncılık.

Sipahi, B., Yurtkoru, E. S. ve Çinko, M. (2010). *Sosyal bilimlerde SPPS'le veri analizi* (3. baskı). İstanbul: Beta Yayıncılık.

Smith, M. J. (1998). "Hayır" *dediğimde kendimi suçlu hissediyorum* (G. Güvenç, Çev.). Ankara: Hekimler Yayın Birliği.

Szpalski, M., Gunzburg, R., & De Kleuver, M. (2003). Unethical research funding contracts: Just say no! *European Spine Journal, 12*(2), 107-107. doi:10.1007/ s00586-003-0551-x

Şimşek, Ö. F. (2007). *Yapısal eşitlik modellemesine giriş: Temel ilkeler ve LISREL uygulamaları*. Ankara: Ekinoks Yayınları.

Tokur Kesgin, M. (2012). Protection of children from hazards of smoking: Community health nurse and primary responsibilities. *Hacettepe University Faculty of Health Sciences Nursing Journal, 19*(1), 90-96.

Tuna Özcivanoğlu, M. E. (2010). Structured group counseling program: Assertiveness skills. *Abant İzzet Baysal University Journal of Faculty of Education, 10*(1), 11-19.

Vaz, S., Parsons, R., Passmore, A. E., Andreou, P., & Falkmer, T. (2013). Internal consistency, test-retest reliability and measurement error of the self-report version of the social skills rating system in a sample of Australian adolescents. *Plos One, 8*(9), 1-8. doi:10.1371/journal.pone.0073924

GENİŞ ÖZET

Giriş

Her bireyin, kişisel, siyasal, sosyal ve ekonomik hakları dışında içinde bulundukları durumdan kaynaklanan başka hakları da bulunmaktadır. Yazılı olmayıp yasalarla da güvence altına alınmamış olan bu haklara "atılganlık hakları" adı verilmektedir (Garner, 2012). "Hayır" diyebilme de birer atılganlık hakkı olarak kabul edilmektedir (Pfeiffer, 2010). "Hayır" diyebilme, bir atılganlık hakkı olsa da her birey tarafından sergilenememektedir. Bu durumun, "hayır" diyebilmenin bir beceri olmasıyla ilgili olduğu düşünülmektedir. Söz konusu beceri, mevcut çalışmada, bireyin kişisel hak ve sınırlarının ihlal edildiği ya da ihlal edilme ihtimalinin bulunduğu durumlarda kendisine yöneltilen talep, teklif ve davranışları, "hayır" diyerek reddedebilmesi ve potansiyel güdümleme çabaları karşısında direnebilmesi şeklinde ele alınmaktadır. "Hayır" diyebilme becerisinin bu şekilde ele alınması, ilgili becerinin hayata geçirilmesi konusunda bir karar alındıktan sonra bu kararın reddetme ve direnme aşamalarında eyleme dökülmesi gerekliliğinden kaynaklanmaktadır.

Reddetme, ilgili zamanda yapılmak istenmeyen bir şeye verilen doğal bir tepkiye işaret etmektedir (Bragger, 1982). "Hayır" diyebilmeye ilişkin reddetme aşamasında bireyden, en uygun sözel ifadeleri ve bunlarla uyumlu vücut dilini kullanarak ilgili talebi, "hayır" deyip kabul etmediğini belirtmesi istenmektedir. Ancak bireyin reddetme davranışı, Kersey ve Masterson'ın (2011) da belirttiği gibi karşı tarafın ısrar gibi çeşitli güdümleme çabalarına neden olabilmektedir. Bu yüzden direnme aşamasında bireyden, karşı taraftan gelebilecek olan güdümleme çabaları karşısında, "hayır" deme konusunda aldığı karardan vazgeçmemesi beklenmektedir.

Reddetme ve direnme aşamaları çerçevesinde eyleme dönüştürülen "hayır" diyebilme becerisi, özellikle çocukları madde kullanmak (Tokur Kesgin, 2012), cinsel açıdan istismar edilmek (Elliott, Browne ve Kilcoyne, 1995; Lecler, Wortley ve Smallbone, 2011), zaman yönetimi ile ilgili sorun yaşamak (Mackenzie ve Nickerson, 2009), ahlaki ihlallerde bulunmak (Leming, 1997; Szpalski, Gunzburg ve De Kleuver, 2003) ve çevrimiçi ortamların risklerine maruz kalmak (Bal ve Kahraman, 2015) gibi sorunlardan koruyabileceği için önem arz etmektedir. Çünkü bu becerilerin erken yaşlarda geliştirilmesi gerekmektedir. 1-3. sınıflar, bu açıdan uygun dönemler olarak değerlendirilirken 4 ve 6. sınıflar, bu becerilerin geliştirilmesi açısından daha fazla geç kalınmaması gereken dönemler olarak düşünülmektedir (Herrmann ve McWhirter, 1997). Ayrıca bu konuda erken adımlar atılması, bu becerilerin daha çok önem kazandığı ileriki gelişim dönemleri açısından kolaylaştırıcı bir zemin hazırlayabilmektedir (Belgrave, Reed, Plybon ve Corneille, 2004; Scheier, Botvin, Diaz ve Griffin, 1999). Ancak çocuklar açısından böyle bir zeminin hazırlanıp hazırlanmadığına yönelik alan yazın incelendiğinde çocukların "hayır" diyebilme becerilerini doğrudan ve ayrıntılı bir biçimde ele alan

çalışmalara rastlanmamaktadır. Bunun yerine söz konusu beceri, çeşitli çalışmalarda (Durualp ve Aral, 2010; Gündoğdu, 2012; Tuna Özcivanoğlu, 2010) yetişkinlerin atılganlık becerileri kapsamında sınırlı bir biçimde ele alınmaktadır. Bunun temel nedenlerinden birinin, bu konuda herhangi bir araştırma yürütülmesini kolaylaştıracak veri toplama araçlarının bulunmaması olduğu söylenebilmektedir. Bu yüzden mevcut araştırmada çocukların "hayır" diyebilme becerilerini değerlendirmeye yönelik bir ölçeğin geliştirilmesi amaçlanmaktadır.

Yöntem

Bu çalışma, bir ölçek geliştirme çalışmasıdır. Bu çalışmanın katılımcılarını, HÖ'nün kapsam geçerliğini sağlamak üzere HÖ'de yer alan maddeler hakkında görüşü alınan uzmanlar ile HÖ'nün yapı geçerliği, ölçüm güvenirliği ve madde ayırtediciliği hakkında fikir edinmek için HÖ'de yer alan maddeleri puanlaması istenen ilkokul 4. sınıf öğrencileri oluşturmaktadır. Bu araştırmada veri toplama aracı olarak sırasıyla HÖ'ye ait 24 maddelik 3'lü Likert, 22 maddelik 3'lü Likert, 12 maddelik 3'lü Likert ve 12 maddelik 5'li Likert olmak üzere dört ayrı taslak form kullanılmıştır. HÖ'nün geliştirilmesi sürecinde öncelikle bu ölçekle ölçülmesi planlanan reddetme ve direnme boyutları yapısal olarak tanımlanmıştır. Daha sonra sırasıyla ölçekte yer alması planlanan maddeler, kapsam geçerliği açısından, Davis (1992) tekniği çerçevesinde, uzman görüşüne sunulmuştur. Ölçeğin yapı geçerliği Açımlayıcı ve Doğrulayıcı Faktör Analizleri ile test edilmiştir. Ölçüm geçerliği hakkında yorum yapabilmek için Cronbach Alfa ve test yarılama (eşdeğer yarılar) yöntemlerine başvurulmuştur. Maddelerin ayırt edicilik düzeyleri ise Düzeltmiş Madde Toplam Korelasyonları ve alt ölçeklere ilişkin ortalama puanlarına göre belirlenmiş en üst ve en alt düzeyde bulunan % 27'lik grupların her bir maddeye ilişkin puanları arasında anlamlı farklılık olup olmadığını belirlemek için uygulanan Bağımsız Gruplar için t-testlerinden yola çıkılarak hesaplanmıştır.

Sonuç ve Tartışma

HÖ, ilkokul 4. sınıf öğrencilerinin "hayır" diyebilme becerilerini ölçmeye dönük öz-bildirimli bir ölçektir. Bu ölçek, reddetme ve direnme boyutlarından oluşmaktadır. Reddetme ve direnme boyutları, 6'şar madde içermektedir. Öğrencilerden maddelerin içerdiği her ifadeyi "Hiçbir zaman", "Nadiren", "Bazen", "Çoğu zaman" ve "Her zaman" şeklinde 5'li Likert bir derecelendirme ile puanlamaları istenmektedir. HÖ'den elde edilen puanlar değerlendirilirken reddetme boyutu için yapılan kodlamalar aynen; direnme boyutu için yapılan kodlamalar ise ters puanlanmaktadır.

Bu araştırmadan elde edilen bulgular, HÖ'nün ilkokul 4. Sınıf öğrencilerinin "hayır" diyebilme becerilerini ölçmek için geçerli bir ölçme aracı olduğuna işaret etmektedir. Araştırma kapsamında işe koşulan güvenirlik testleri, HÖ ile güvenilir ölçümler yapılabileceğini göstermektedir. Madde analizleri ise HÖ'ye yanıt veren bireylerin, "hayır" diyebilme performansları açısından ayırt edilebileceklerini ortaya koymaktadır. Bu sonuçlar, genel olarak çocukların "hayır" diyebilme becerileri ile ilgili yapılacak çalışmalarda HÖ'nün geçerli ve güvenilir sonuçlar verebilecek bir ölçme aracı olarak kullanılabileceğini kanıtlamaktadır.

"Hayır" diyebilme ile ilgili diğer ölçme araçları (García-Ros, Pérez-González, & Hinojosa, 2004; Macan, Shahani, Dipboye, & Philips, 1990; Scheier vd., 1999) incelendiğinde, bu ölçme araçlarının "hayır" diyebilme becerisini farklı amaçlarla geliştirilmiş ölçeklerde, sınırlı bir biçimde, madde (item) düzeyinde içerdiği ya da bu beceriyi sadece madde kullanımı ve zaman yönetimi gibi bağlamlarda değerlendirmeyi amaçladığı anlaşılmaktadır. Mevcut araştırma ise söz konusu beceriyi, reddetme ve direnme boyutlarında ayrıntılı bir biçimde ölçebilecek ve birçok bağlamda kullanılabilecek bir ölçme aracıyla alan yazına katkı sağlamayı amaçlamaktadır.

Bu araştırmanın bazı sınırlılıkları bulunmaktadır. Bu sınırlılıklar çerçevesinde araştırma verileri, sadece 9-11 yaş aralığındaki ilkokul 4. sınıf öğrencilerinden toplanmıştır. Dolayısıyla ilerleyen dönemlerde yapılacak araştırmalarda HÖ'nün farklı yaş grupları ve gelişim dönemlerindeki bireyler için de geçerlik ve güvenirlik çalışmalarının yapılması önerilmektedir. Böylece bireylerin "hayır" diyebilme becerilerinin farklı gelişim dönemleri açısından nasıl bir seyir izlediği ortaya

konulabilecektir. Bu çalışmada HÖ'nün kapsam ve yapı geçerliği incelenmişken uyum geçerliği incelenmemiştir. Gelecekte yapılacak çalışmalarla HÖ'nün uyum geçerliğinin de yapılmasının faydalı olacağına inanılmaktadır. Bu kapsamda "hayır" diyebilme becerisi ile ilişkili olabilecek atılganlık becerilerini ya da sosyal becerileri (Caldarella ve Merrell, 1997; Eslami, Mazaheri, Mostafavi, Abbasi ve Noroozi, 2014; Gresham ve Elliot, 1990; Vaz, Parsons, Passmore, Andreou ve Falkmer, 2013) ölçmeye yönelik geliştirilmiş olan veri toplama araçlarının işe koşulması önerilmektedir. Mevcut çalışmada HÖ'nün ölçüm güvenirliği, iç tutarlılık katsayısının belirlenmesi ve test yarılama (eşdeğer yarılar) yöntemleriyle tespit edilmiştir. Yapılacak farklı çalışmalarla HÖ'nün ölçüm güvenirliği konusunda test-tekrar test yönteminin de kullanılabileceği düşünülmektedir. Bu şekilde ölçeğin farklı zamanlarda tutarlı ölçümler sunup sunamayacağı anlaşılabilecektir.

Appendices

Appendix 1: The Scale of Being Able to Say "No"[*]

FORM I Dear students, • Read each sentence below carefully. • Choose the best alternative for you. • Do not leave any question unanswered. • Choose only one alternative for each item.	**Never**	**Rarely**	**Sometimes**	**Mostly**	**Always**
1. I can say "**No; I don't want to do it**" when somebody asks me to do something that I don't like doing.	O	O	O	O	O
2. I can say "**No; you cannot do this to me**" when somebody does something that I don't like to me.	O	O	O	O	O
3. I can say "**No; I will not do this**" when somebody asks me to do something that will cause me trouble.	O	O	O	O	O
4. I can say "No." **unashamedly** when I don't want to do something.	O	O	O	O	O
5. I can say "No." **to suggestions that I may regret accepting**".	O	O	O	O	O
6. I can say "No." **to suggestions that I can be angry** with myself if I accept.	O	O	O	O	O
7. When somebody to whom I said "no" comes to me **with the same demand** after a while, I do what he/she asks me to do by giving up my decision.	O	O	O	O	O
8. When somebody to whom I said "no" **accuses me**, I do what he/she asks me to do by giving up my decision.	O	O	O	O	O
9. When somebody to whom I said "no" **cries**, I do what he/she asks me to do by giving up my decision.	O	O	O	O	O
10. When I see somebody to whom I said "no" **to be sad**, I do what he/she asks me to do by giving up my decision.	O	O	O	O	O
11. When somebody to whom I said "no" **pulls some strings** I do what he/she asks me to do by giving up my decision.	O	O	O	O	O
12. When somebody to whom I said "no" **is cross with me**, I do what he/she asks me to do by giving up my decision.	O	O	O	O	O

[*] The English version of the scale is shared only in terms of providing information about the items in the Turkish version. The validity and reliability study of this form or the adaptation study from Turkish to English was not conducted.

Appendix 2: The Original Scale of Being Able to Say "No"

I.FORM Sevgili öğrenciler, • Lütfen, aşağıdaki her cümleyi ayrı ayrı okuyun. • Sizin için hangi seçenek uygunsa onu işaretleyin. • Hiçbir maddeyi boş bırakmayın. • Her madde için sadece bir seçeneği işaretleyin.	**Hiçbir zaman**	**Nadiren**	**Bazen**	**Çoğu zaman**	**Her zaman**
1. Biri, benden yapmaktan hoşlanmadığım bir şey istediğinde ona, "**Hayır, bunu yapmak istemiyorum.**" diyebiliyorum.	O	O	O	O	O
2. Biri, bana hoşuma gitmeyen bir şey yaptığında ona, "**Hayır, bana bunu yapamazsın**." diyebiliyorum.	O	O	O	O	O
3. Biri benden başıma kötü işler açacak bir şey istediğinde ona, **"Hayır, bunu yapmayacağım"** diyebiliyorum.	O	O	O	O	O
4. Bir şeyi yapmak istemediğim zaman, buna **çekinmeden** "Hayır." diyebiliyorum.	O	O	O	O	O
5. Kabul edersem **pişman olabileceğim tekliflere** "Hayır." diyebiliyorum.	O	O	O	O	O
6. Kabul edersem **kendime kızabileceğim tekliflere** "Hayır." diyebiliyorum.	O	O	O	O	O
7. "Hayır" dediğim bir insan, bir süre sonra **benden aynı şeyi tekrar istediğinde** bu kararımdan vazgeçip onun benden istediği şeyi yapıyorum.	O	O	O	O	O
8. "Hayır" dediğim bir insan, **beni suçladığında** bu kararımdan vazgeçip onun benden istediği şeyi yapıyorum.	O	O	O	O	O
9. "Hayır" dediğim bir insan, **ağladığında** bu kararımdan vazgeçip onun benden istediği şeyi yapıyorum.	O	O	O	O	O
10. "Hayır" dediğim **bir insanın üzüldüğünü gördüğümde** bu kararımdan vazgeçip onun benden istediği şeyi yapıyorum.	O	O	O	O	O
11. "Hayır" dediğim bir insan, **araya sevdiğim birini koyduğunda** bu kararımdan vazgeçip onun benden istediği şeyi yapıyorum.	O	O	O	O	O
12. "Hayır" dediğim bir insan, **bana küstüğünde** bu kararımdan vazgeçip onun benden istediği şeyi yapıyorum.	O	O	O	O	O

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);49-63

Kategori Sayısının Psikometrik Özellikler Üzerine Etkisinin Mokken Homojenlik Modeli'ne Göre İncelenmesi*

Investigation of the Effects of the Number of Categories on Psychometric Properties According to Mokken Homogeneity Model*

Asiye ŞENGÜL AVŞAR**

Öz

Araştırmanın amacı çok kategorili puanlanan maddelerden oluşan testlerde kategori sayısının psikometrik özellikler üzerindeki etkisinin parametrik olmayan madde tepki kuramı (POMTK) modeli ile belirlenmesidir. Belirlenen amaç doğrultusunda iki farklı büyüklükte (100 ve 500), çeşitli dağılım özelliklerine sahip (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan) örneklemler için iki farklı test uzunluğunda (10 madde ve 30 madde), üç farklı sayıda kategoriye (üç, beş ve yedi) sahip maddeler simülatif olarak üretilmiştir. Kategori sayısının psikometrik özellikler üzerindeki etkisi POMTK modellerinden Mokken Homojenlik Modeli (MHM) ile araştırılmıştır. Yapılan araştırma temel araştırma olarak tasarlanmıştır. Verilerin üretilmesinde ve verilerin analizinde R Studio 3.4.0 yazılımı kullanılmıştır. R Studio yazılımında MHM'ye göre analizler Mokken paketi ile yapılmıştır. MHM'ye göre yapılan ölçekleme sonucunda kategori sayısının değişmesiyle birlikte maddelerin MHM'ye uyumunda belli bir örüntü gözlenmemiştir. Genel olarak hem kısa testlerde, hem de uzun testlerde kategori sayısının güvenirlik değerlerinin kestiriminde etkili olmadıkları gözlenmiştir. Araştırmada belirlenen test koşullarında testler MHM'ye düşük düzeyde uyumlu çıkmıştır.

Anahtar Kelimeler: çok kategorili puanlanan maddeler, kategori sayısı, parametrik olmayan madde tepki kuramı, mokken homojenlik modeli

Abstract

The aim of the research was to examine the effects of the number of categories for polytomous items on psychometric properties in a nonparametric item response theory (NIRT) model. For the purpose of the study, data sets with two different sample sizes (100 and 500) that come from different sample distribution shapes (normal distribution, positively skewed distribution, and negatively skewed distribution), two different test lengths (10 items and 30 items), and three different number of categories (three, five, and seven) were generated. The effects of the number of categories on psychometric properties of polytomous items were analyzed by Mokken Homogeneity Model (MHM) under NIRT model. The research was designed as a basic research. In the generation and analysis of data sets, R Studio 3.4.0 software was used. For analysis conducted with MHM, Mokken package was used in R Studio. According to scaling with MHM, specific pattern of item fit to MHM with changing the number of categories was not observed. In general, it was found that the number of categories has no effect on reliability estimate. It was determined that tests have weak fit to MHM under test conditions in the research.

Keywords: polytomous items, number of category, nonparametric item response theory, mokken homogeneity model

* Bu araştırma 1-3 Eylül 2016 tarihinde Akdeniz Üniversitesi'nde düzenlenen V. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi'nde sözlü olarak sunulan bildiriden türetilmiştir.

** Dr. Öğretim Üyesi, Recep Tayyip Erdoğan Üniversitesi, Eğitm Fakültesi, Rize-Türkiye, asiye.sengul@erdogan.edu.tr, ORCID ID: orcid.org/0000-0001-5522-2514

DOI: 10.21031/epod.357160
Geliş Tarihi: 23.11.2017
Kabul Tarihi: 29.01.2018

GİRİŞ

Bireylerin duyuşsal özelliklerinin ölçülmesi, eğitimde ve psikolojide önemli yer tutmaktadır. Bu özellikler ölçülürken kullanılan ölçme araçlarının ikili puanlanan maddelerden daha çok tepkilerin dereceli ya da çoklu olarak sunulduğu yanıt kategorilerine sahip maddelerden oluştukları görülmektedir. Çok kategorili puanlanan maddeler, ikili puanlanan maddelerle karşılaştırıldığında tepki kategorilerinin sayısı daha fazla olduğundan ilgilenilen örtük özelliği daha geniş ranjlarda ölçebilmekte ve buna bağlı olarak örtük özellikle ilgili daha fazla bilgiye ulaşılmasını sağlamaktadır (Ostini ve Nering, 2006).

Ölçme araçlarının psikometrik özelliklerinin belirlenmesi, bu araçlara göre verilen kararların doğruluğu ve uygunluğu açısından çok önemlidir. Psikometrik özellikler, ölçmeleri etkileyen problemleri araştıran ve bu problemleri mümkün olduğunca azaltmaya çalışan bir disiplin olarak tanımlanan test kuramları ile belirlenir (Crocker ve Algina, 1986). Eğitimde ve psikolojide sıklıkla Klasik Test Kuramı'na (KTK) ve Madde Tepki Kuramı'na (MTK) göre ölçekleme yapılarak ölçme araçlarının psikometrik özelliklerinin belirlendiği görülmektedir.

Genellikle istatistiksel bir model olarak tanımlanan MTK, bireyin madde ve test performansını test eden ve bireyin performansının altında yatan yeteneğini, maddeler ve test aracılığıyla kestiren bir kuram olarak tanımlanabilir (Hambleton ve Jones, 1993). MTK literatürde son zamanlarda yapılan çalışmaların katkısıyla genel olarak parametrik madde tepki kuramı (PMTK) ve parametrik olmayan MTK (POMTK) (Sijtsma ve Molenaar, 2002) modelleri olacak şekilde genel iki başlık altında sınıflandırılmıştır.

PMTK modellerine göre yeni sayılabilen POMTK modelleri, PMTK modelleri ile karşılaştırıldığında daha az varsayım gerektirmektedir (Štochl, 2007). POMTK modelleri kısa testlerde, küçük örneklemlerde uygulamalarda kolaylık sağlayan modellerdir (Junker ve Sijtsma, 2001; Meijer, 2004; Molenaar, 2001). Literatür incelendiğinde POMTK modellerinin Mokken model ve parametrik olmayan regresyon modelleri olacak şekilde sınıflandırılabileceği görülmektedir (Şengül Avşar ve Tavşancıl, 2017). Mokken model; Monoton Homojenlik Modeli (MHM) ve İkili Monotonluk Modeli (İMM) olacak şekilde kendi için alt modellere ayrılmaktadır (Sijtsma ve Molenaar, 2002).

MHM sıralama düzeyindeki ölçmelerde, bireylerin sıralama amacıyla değerlendirilmesinde kullanılan (Štochl, 2007) ve bir testi alan bireylerin puanlarını kullanarak onları örtük özellikleri boyunca sıralayan (Tendeiro ve Meijer, 2013) bir model olarak tanımlanabilir. MHM hem ikili (1-0) puanlanan maddelerden oluşan ölçme araçlarının hem de çok kategorili puanlanan maddelerden oluşan ölçme araçlarının POMTK'ya göre ölçeklenmesini sağlayan bir modeldir. MHM, çok kategorili puanlanan maddeler için PMTK modellerinden dereceli tepki modelinin (DTM-graded response model) parametrik olmayan karşılığı olarak tanımlanmaktadır (Hemker, Sijtsma, Molenaar ve Junker, 1996; Sijtsma ve Molenaar, 2002).

Ölçme araçlarının psikometrik özelliklerinin MHM'ye göre belirlenmesinde hem ikili puanlanan maddeler hem de çok kategorili puanlanan maddeler için MHM'ye göre parametre kestirimleri, "ölçeklenebilirlik katsayısı (scalability coefficient-H)" ile yapılmaktadır (van Onna, 2004). İkili puanlanan maddeler için Loevinger (1947,1948) tarafından geliştirilen H katsayısı; Mokken (1971) tarafından MHM'de bir set içinde yer alan madde çiftleri (i, j madde çiftleri) için (H_{ij}), tek bir maddenin setteki diğer maddelerle (H_i) ve madde setlerinin tamamıyla (H) olan ilişkisini tanımlanmak için yeniden düzenlenmiştir (Mokken, 1997).

Mokken modelleriyle yapılan ölçeklemede çok önemli bir yere sahip olan H katsayısı iki veya üç parametreli lojistik modellerde yer alan "a" katsayısının (madde ayırt edicilik indeksi) parametrik olmayan karşılığı olarak yorumlanabileceği gibi ölçme araçlarının MHM'ye göre ölçeklenip ölçeklenmediğinin belirlenmesinde kullanılan ölçeklenebilirlik indeksi anlamlarını taşımaktadır (Meijer, 2004; Mokken, 1997; van Onna, 2004).

Mokken modellerinde güvenirlik hesaplamaları için Cronbach α, Guttman lambda 2 (λ) ve Rho katsayılarının raporlaştırıldığı görülmektedir (Şengül Avşar ve Tavşancıl, 2017). Mokken (1971)

tarafından önerilen ve literatürde aynı zamanda Molenaar Sijtsma (MS) istatistiği olarak bilinen Rho katsayısı Mokken modellerinden İMM'ye uygun bir güvenirlik katsayısıdır (Štochl, 2007; van der Ark, 2015). MHM'ye göre ölçeklenen ölçme araçlarının güvenirlik kestirimleri için ayrıca van der Ark, van der Palm ve Sijstma'nın (2011) geliştirdikleri örtük sınıf güvenirlik katsayısı (LCRC-latent class reliability coefficient) kullanılabilir.

POMTK modellerinden; tek boyutluluk, yerel bağımsızlık ve monotonluk varsayımlarını gerektiren MHM'nin (Sijtsma ve Molenaar, 2002) örtük değişkenle, homojen ve monoton madde karakteristik eğrilere (MKE) sahip maddeler arasındaki ilişkiyi tanımlayan bir model olduğu ifade edilmektedir (Meijer ve Baneke, 2004). Burada MHM'nin tek boyutlu PMTK modelleriyle benzer varsayımları gerektiği açık bir şekilde görülmektedir. Ancak PMTK ve POMTK modelleri karşılaştırıldığında iki model arasındaki temel farkın ikili puanlanan maddeler için MKE'lere, çok kategorili puanlanan maddeler için madde adım fonksiyonlarına bağlı olduğu bilinmektedir (Şengül Avşar ve Tavşancıl, 2017). MKE'ler PMTK modellerinde monoton ve lojistik olarak kestirilirken bu eğriler POMTK modellerinde monoton olmalarına rağmen lojistik olarak kestirilmezler (Sijtsma ve Molenaar, 2002).

Literatür taramasında yurt içinde ve yurt dışında ölçeklerin psikometrik özelliklerinin POMTK modellerine göre incelendiği çeşitli çalışmaların yapıldığı görülmüştür (Galindo Garre ve diğerleri, 2014; Sachs, Law ve Chan, 2003; Koğar, 2015; Şengül Avşar ve Tavşancıl, 2017; Young, Blodgett ve Reardon, 2003). Özellikle çok kategorili puanlanan maddeleri konu alan simülatif araştırmalarda örneklem büyüklüğü, madde sayısı, örneklemin dağılım şekli gibi çeşitli faktörlerin POMTK modellerinden MHM'ye göre elde edilen geçerlik ve güvenirlik katsayılarına etkisi araştırılırken gerçek veri setleriyle yapılan araştırmalarda ölçeklerin psikometrik özelliklerinin belirlenmesi amaçlanmıştır.

Duyuşsal özelliklerin ölçülmesinde Likert tipi ölçekler sıklıkla kullanılmaktadır. Bu ölçekler sıralama düzeyinde, çok kategorili puanlanan maddelerden oluşmaktadır. Literatürde Likert tipi, çok kategorili puanlanan maddelerden oluşan ölçme araçlarının psikometrik özelliklerini etkileyen önemli bir faktörün kategori sayısı olduğu ifade edilmektedir (Fabiola, Iwin, Jennifer ve Zaira, 2012; Leung, 2011; Lozano, García-Cueto, ve Muñiz, 2008; Maydeu-Olivares, Kramp, García-Forero, Gallardo-Pujol ve Coffman, 2009; Preston ve Colman, 2000; Weng, 2004).

Weng (2004) araştırmasında kategori sayısının psikometrik özellikler üzerindeki etkisinin literatürde farklı araştırmacılar tarafından pek çok kez araştırıldığını, bazı araştırmacıların özellikle iç tutarlılık katsayısı olarak yorumlanan Cronbach α üzerindeki etkilerinin önemli olduğunu belirtmiştir. Bununla birlikte Maydeu-Olivares ve diğerleri (2009) araştırmalarında ideal kategori sayısının ne olması gerektiği konusunda literatürde fikir birliği olmadığını ancak kategori sayısının ölçeklerin psikometrik özellikleri üzeride etkileri olduğunu belirtmiştir.

Yapılan literatür taramasında KTK kapsamında kategori sayının psikometrik özellikler üzerindeki etkisinin araştırıldığı çeşitli çalışmaların yapıldığı görülmektedir (Erkuş, Sanlı, Bağlı ve Güven, 2000; Leung, 2011; Lozano ve diğerleri, 2008; Maydeu-Olivares ve diğerleri, 2009; Preston ve Colman, 2000; Weng, 2004; Uyumaz ve Çokluk, 2016). Bu araştırmalarda genel olarak kategori sayısının güvenirlik kestirimleri üzerinde etkili olduğu, kategori sayısı arttıkça güvenirlik değerlerinin arttığı sonucuna ulaşılmıştır.

Kategori sayısının psikometrik özellikler üzerindeki etkisini PMTK modelleriyle araştıran çeşitli çalışmalar da bulunmaktadır (Fabiola ve diğerleri, 2012; İlhan ve Güler, 2017; Lee ve Paek, 2014; Maydeu-Olivares ve diğerleri, 2009). Fabiola ve diğerleri (2012) ve Maydeu-Olivares ve diğerleri (2009) kategori sayısının psikometrik özellikler üzerinde etkili olduğunu ifade ederken Lee ve Paek (2014) kategori sayısının psikometrik özellikler üzerinde etkili olmadığını ifade etmiştir. Ayrıca İlhan ve Güler (2017) yaptıkları araştırmada kategori sayısının model veri uyumu üzerinde önemli bir etkisinin olmadığını belirtmişlerdir.

Yapılan literatür taramasında küçük örneklemlere uygulanan ölçme araçlarının psikometrik özelliklerinin belirlenmesinde PMTK modellerinin oldukça sınırlayıcı olduğu, bu gibi durumlarda POMTK modellerinden MHM'nin PMTK modellerine göre kullanışlı bir model olduğu ve

uygulamalarda kolaylık sağlayabileceği ifade edilmiştir. Çok kategorili puanlanan maddelerden oluşan ölçme araçlarının psikometrik özelliklerinin incelenmesinde kategori sayısının önemli bir faktör olduğu belirtilmiştir. Yapılan bu araştırmada farklı kategori sayılarına sahip ölçme araçlarının psikometrik özelliklerinin simülatif veriler aracılığıyla uygulamalarda kolaylık sağlayan POMTK modellerinden MHM ile belirlenmesi gerekli görülmüştür.

Araştırmanın Amacı

Bu araştırmanın genel amacı, simülatif olarak üretilen "üç", "beş" ve "yedi" kategoriden oluşan 10 maddelik kısa ve 30 maddelik uzun testlerin, çeşitli dağılım özelliklerine (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan) sahip 100 ve 500 kişilik örneklemlerden oluşan test koşullarında psikometrik özelliklerinin MHM'ye göre belirlenmesidir. Belirlenen genel amaç doğrultusunda aşağıdaki sorulara yanıt aranmıştır.

Çeşitli özelliklere sahip örneklemelere uygulanan kategori sayısı ""üç", "beş" ve "yedi" olan 10 maddelik kısa ve 30 maddelik uzun testlerin MHM'ye göre elde edilen:

1. Maddelere ait model veri uyumu düzeyleri nasıldır?
2. Maddelere ait model veri uyumu değerleri için kestirilen standart hata değerleri nelerdir?
3. Testlere ait model veri uyumu değerleri nelerdir?
4. Testlere ait model veri uyumu değerleri için kestirilen standart hata değerleri nelerdir?
5. Testler için kestirilen güvenirlik değerleri (α, λ ve LCRC) nelerdir?

YÖNTEM

Bu araştırma; kategori sayısının çok kategorili puanlanan maddelerden oluşan testlerin psikometrik özellikleri üzerine etkilerinin simülatif olarak üretilen veriler aracılığıyla belirlenmesinin amaçlandığı temel bir araştırmadır. Temel araştırmalarda mevcut teorilere dayanılarak yeni bilgi ya da yeni teoriler üretilir (Freankal, Wallen ve Hyun, 2012).

Araştırmanın test koşulları Tablo 1'de verilmiştir. Araştırmada yer alan koşullar ve verilerin üretilmesi için gerekli parametrelerin seçimi, literatür taraması sonuçlarına göre daha önce yapılan benzer çalışmalar incelenerek belirlenmiştir.

Tablo 1. Test Koşulları

		Madde Sayısı					
		10 Madde			30 Madde		
Örneklemin Dağılım Şekli	Örneklem Büyüklüğü	3 kategori	5 kategori	7 kategori	3 kategori	5 kategori	7 kategori
Normal Dağılım	100	X	X	X	X	X	X
	500	X	X	X	X	X	X
Sağa Çarpık Dağılım	100	X	X	X	X	X	X
	500	X	X	X	X	X	X
Sola Çarpık Dağılım	100	X	X	X	X	X	X
	500	X	X	X	X	X	X

Tablo 1'de araştırmada; üç farklı örneklem dağılım şekli, iki farklı örneklem büyüklüğü, iki farklı test uzunluğu ve üç farklı kategori sayısı olmak üzere 36 farklı (3*2*2*3) test koşulunun incelendiği görülmektedir. Tüm bu koşullar için 100 replikasyon yapılarak toplamda 3600 veri seti oluşturulmuştur. Tablo 1'de yer alan tüm test koşullarına literatüre dayalı yapılan incelemeler sonucunda karar verilmiştir.

Örneklem büyüklüğü için POMTK'ya göre yapılacak çözümlemelerde, Molenaar (2001) 300-400 kişilik örnekleme, Ramsay (1991) ise en az 100 kişilik örnekleme ihtiyaç duyulduğunu belirtmiştir. Ayrıca veriler POMTK modellerinden MHM'nin, PMTK'da parametrik karşılığı olan DTM'ye göre üretilmiştir. DTM'ye uyum gösteren veriler MHM'ye de uyum göstermektedir (Emons, 2008). DTM için en az 500 kişiden elde edilecek kestirimlerin doğru olduğu belirtilmektedir (Jiang, Wang ve Weiss, 2016). Literatürdeki bu bilgilere dayalı olarak araştırmada 100 kişiden oluşan küçük örneklemler ve 500 kişiden oluşan büyük örneklemler tercih edilmiştir. Ayrıca araştırmada yetenek dağılımlarının normal, sağa çarpık ve sola çarpık dağıldığı durumlar incelenmiştir. Yetenek dağılımları ortalamaları sıfır, standart sapmaları bir olan normal dağılımdan üretilmiş olup Tablo 2'de yetenek dağılımlarına ilişkin betimsel istatistikler verilmiştir.

Tablo 2. Yetenek Dağılımlarına İlişkin Betimsel İstatistikler

Örneklemin Dağılım Şekli	Örneklem Büyüklüğü	Çarpıklık Katsayısı	Basıklık Katsayı	Standart Hata
Normal Dağılım	100	-0.01	-0.18	0.10
	500	-0.01	-0.06	0.04
Sağa Çarpık Dağılım	100	0.54	0.06	0.10
	500	0.59	0.24	0.04
Sola Çarpık Dağılım	100	-0.55	0.01	0.10
	500	-0.57	0.18	0.04

Çarpıklık katsayısının mutlak değerinin; birden büyük olduğu durumda örneklemlerin dağılım şekillerinin yüksek düzeyde çarpık, 0.50 ve bir arasında orta düzeyde çarpık, 0.50'den küçük olduğu durumlarda ise yaklaşık olarak simetrik olduğu ifade edilmektedir (Bulmer, 1979, s. 63). Tablo 2 incelendiğinde çarpık dağılan örneklemlerin orta düzeyde çarpık olduğu görülmektedir. Eğitimde ve psikolojide uygulamalarda normal dağılımlı veri setlerine ulaşmak hedeflenmektedir. MTK uygulamalarında da mümkün olduğunca veri setinin normal dağılımlı olması beklenir. Bu bağlamda çarpık dağılımların normal dağılımdan aşırı düzeyde sapmış olmaması gerekir. Bunun için gerçek uygulamaları düşünerek araştırmada çarpık dağılımlar orta düzeyde çarpık olacak şekilde seçilmiştir.

Madde parametreleri literatüre bağlı kalınarak normal ve tek biçimli olacak şekilde belirlenmiştir. Buna göre b parametresi $N\sim(0, 1)$ normal, a parametresi ise tek biçimli $a \in U\ [1,2]$ olacak şekilde seçilmiştir (Bahry, 2012; Cohen, Kim ve Baker, 1993; DeMars, 2002; Syu, 2013).

Literatürde test uzunluğunu ya da madde sayısını konu alan araştırmalar incelendiğinde 10'dan 80'e kadar çeşitli sayılarda maddelerden oluşan testlerin kullanıldığı görülmektedir (Lee, 2007; Lee, Wollack ve Douglas, 2009; Liang, Wells ve Hambleton, 2014; Patsula ve Gessaroli, 1995; Stone, 1992; Stone ve Zhang, 2003; Sueiro ve Abad, 2011). Bu araştırmada POMTK modellerinin kısa testlerde uyum göstermesi avantajı düşünülerek az sayıda maddelerden oluşan testler tercih edilmiştir. Buna göre araştırma kapsamında 10 maddelik testler kısa, 30 maddelik testler uzun testler olarak belirlenmiştir.

Araştırmada kategori sayılarının belirlenmesinde KTK ve MTK kapsamında kategori sayısının psikometrik özellikler üzerindeki etkisini araştıran çalışmalardan yararlanılmıştır (Fabiola ve diğerleri, 2012; Leung, 2011; Uyumaz ve Çokluk, 2016; Weng, 2004).

Verilerin Analizi

Araştırmada belirlenen test koşullarına uygun simülatif verilerin üretilmesi ve verilerin analizi R Studio 3.4.0 yazılımı ile gerçekleştirilmiştir. MHM'ye göre yapılan çözümlemeler ve güvenirlik katsayıları hesaplamaları için van der Ark (2007) tarafından geliştirilen Mokken paketi kullanılmıştır.

Araştırmada maddelerin ve testlerin MHM'ye uyumları H katsayıları ile belirlenmiştir. H katsayılarının değerlendirilmesinde Mokken (1971) tarafından tanımlanmış değerlendirme ölçütleri kullanılmaktadır. Bu ölçütler; $0.30 \leq H < 0.40$ için düşük, $0.40 \leq H < 0.50$ için orta ve $H \geq 0.50$ için yüksek olacak şekilde belirlenmiştir (Meijer ve Baneke, 2004; Mokken, 1971, 1997; Sijtsma, Debets ve Molenaar, 1990; van Onna, 2004). Araştırmada ayrıca maddelere ve testlere ait H değerleri için kestirilen standart hata değerleri incelenmiştir.

BULGULAR

Araştırma kapsamında elde edilen bulgular araştırma sorularının sırasına bağlı olarak sunulmuştur.

Maddelere Ait Model Veri Uyumu Düzeyleri

MHM'ye göre ölçeklenen maddelerin model veri uyumu değerleri olan H katsayılarının Mokken'a (1971) göre belirlenen değerlendirme ölçütleriyle incelenmesi sonucunda elde edilen değerlerin uyum düzeyleri 10 madde ve 30 madde için sırasıyla Tablo 3'te ve Tablo 4'te verilmiştir.

Tablo 3. 10 Madde İçin Model Veri Uyumu Düzeyleri

		Normal Dağılan		Sağa Çarpık Dağılan		Sola Çarpık Dağılan	
	Uyum Düzeyi	100	500	100	500	100	500
3 kategori	Düşük	8	8	7	7	7	8
	Orta	-	1	1	-	3	2
	Yüksek	-	-	-	-	-	-
5 kategori	Düşük	7	8	7	2	7	8
	Orta	2	1	1	6	3	2
	Yüksek	-	-	-	2	-	-
7 kategori	Düşük	6	7	4	6	8	8
	Orta	2	1	2	1	2	2
	Yüksek	-	-	-	-	-	-

Tablo 3 incelendiğinde çeşitli dağılım özelliklerine sahip iki farklı büyüklükteki örneklemlerden elde edilen maddelerin MHM'ye göre ölçeklenmesi sonucu elde edilen H değerlerinin uyum düzeylerinin verildiği görülmektedir. Tablo 3'te sola çarpık dağılan örneklemler hariç diğer örneklemlerde MHM'ye uyum göstermeyen maddelerin olduğu anlaşılmaktadır.

Normal dağılan ve sağa çarpık dağılan hem küçük hem de büyük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayısının beş olduğu söylenebilir. Sola çarpık dağılan küçük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayılarının üç ve beş olduğu söylenebilir. Sola çarpık dağılan büyük örneklemlerde kategori sayısının maddelerin MHM'ye uyum düzeyini etkilemedikleri görülmektedir. Araştırma kapsamında 10 maddelik testler için genel olarak maddelerin MHM'ye en iyi uyum gösterdiği kategori sayısının beş olduğu sonucuna ulaşılabilir.

Tablo 4. 30 Madde İçin Model Veri Uyumu Düzeyleri

		Normal Dağılan		Sağa Çarpık Dağılan		Sola Çarpık Dağılan	
	Uyum Düzeyi	100	500	100	500	100	500
3 kategori	Düşük	20	20	20	20	22	23
	Orta	7	6	4	3	7	7
	Yüksek	-	1	1	1	-	-
5 kategori	Düşük	23	23	21	19	22	23
	Orta	6	6	5	6	7	7
	Yüksek	-	-	-	-	-	-
7 kategori	Düşük	22	22	19	20	22	22
	Orta	7	7	7	6	6	7
	Yüksek	-	-	-	-	-	-

Tablo 4 incelendiğinde çeşitli dağılım özelliklerine sahip iki farklı büyüklükteki örneklemlerden elde edilen maddelerin MHM'ye göre ölçeklenmesi sonucu elde edilen H değerlerinin uyum düzeylerinin verildiği görülmektedir. Tablo 4'te sola çarpık dağılan büyük örneklemlerde üç ve beş kategoriden oluşan test koşulları hariç diğer tüm test koşullarında MHM'ye uyum göstermeyen maddelerin olduğu anlaşılmaktadır.

Normal dağılan hem küçük hem de büyük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayıları beş ve yedidir. Sağa çarpık dağılan örneklemlerde genel olarak kategori sayısı arttıkça maddelerin MHM'ye uyum düzeyleri artmıştır. Bu koşulda MHM'ye en iyi uyum kategori sayısının yedi olduğu görülse de beş ve yedi kategorili puanlanan maddeler arasında MHM'ye uyum düzeyi açısından farkın tek bir madde için olduğu belirtilmelidir. Sola çarpık dağılan hem küçük hem de büyük örneklemlerde maddelerin MHM'ye en iyi uyum gösterdiği kategori sayılarının üç ve beş kategori olduğu görülse de yedi kategorili puanlanan maddelerle bu maddeler arasında MHM'ye uyum düzeyi açısından farkın tek bir madde için olduğu belirtilmelidir.

Araştırma kapsamında 30 maddelik testler için tüm test koşulları birlikte değerlendirildiğinde, genel olarak MHM'ye en iyi uyumların kategori sayısının beş olduğu durumlarda oluştuğu sonucuna ulaşılabilir. Bununla birlikte örneklemlerin dağılım şekilleri kendi içlerinde incelendiğinde sola çarpık dağılan örneklemler hariç diğer durumlarda genel olarak kategori sayısı arttıkça maddelerin MHM'ye uyum düzeylerinin arttığı da söylenebilir.

Araştırmada belirlenen test koşullarında (farklı dağılım özelliklerine sahip farklı büyüklükteki örneklemlerden 10 maddelik kısa ve 30 maddelik uzun testler için) genel olarak kategori sayısının maddelerin MHM'ye uyum düzeyleri üzerinde çok etkili bir faktör olmadığı sonucuna ulaşılmıştır. Her ne kadar araştırmada belirlenen test koşullarının çoğunda kategori sayısı arttıkça maddelerin MHM'ye uyumu artmış gibi görünse de bu artışın tek madde ile sınırlı olduğu belirtilmelidir. Genel olarak tüm test koşulları birlikte değerlendirildiğinde MHM'ye en iyi uyuma beş kategorili puanlanan maddelerden ulaşıldığı ifade edilebilir.

Maddelere Ait Model Veri Uyumu Değerleri İçin Kestirilen Standart Hata Değerleri

Tablo 5'te farklı dağılım özelliklerine sahip farklı büyüklükteki örneklemlerden, 10 maddelik kısa ve 30 maddelik uzun testlerden elde edilen tüm maddelere ait model veri uyumu değerleri için kestirilen standart hata değerlerinin en küçük (SH_{iek}) ve en büyük (SH_{ieb}) değerleri verilmiştir.

Tablo 5. Maddelere Ait Model Veri Uyumu Değerleri İçin Kestirilen En Küçük ve En Büyük Standart Hata Değerleri

		Normal Dağılan				Sağa Çarpık Dağılan				Sola Çarpık Dağılan			
		100 kişi		500 kişi		100 kişi		500 kişi		100 kişi		500 kişi	
		SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}	SH_{iek}	SH_{ieb}
10 madde	3	0.06	0.10	0.03	0.04	0.06	0.09	0.03	0.04	0.06	0.10	0.03	0.05
	5	0.06	0.10	0.03	0.05	0.06	0.09	0.03	0.05	0.06	0.10	0.03	0.05
	7	0.07	0.08	0.03	0.05	0.06	0.07	0.02	0.04	0.06	0.10	0.03	0.05
30 madde	3	0.05	0.08	0.02	0.05	0.05	0.10	0.02	0.05	0.05	0.10	0.02	0.05
	5	0.05	0.09	0.02	0.05	0.05	0.10	0.02	0.05	0.05	0.09	0.02	0.04
	7	0.05	0.09	0.02	0.05	0.05	0.11	0.02	0.05	0.05	0.09	0.02	0.04

Tablo 5'te MHM'ye göre 10 madde için elde edilen maddelere ait SH değerleri; normal dağılan küçük örneklemlerde en küçük 0.06, en büyük 0.10 değerlerini alırken normal dağılan büyük örneklemlerde en küçük 0.03, en büyük 0.05 değerlerini almıştır. Sağa çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.06, en büyük 0.09 değerlerini alırken sağa çarpık dağılan büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Sola çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.06, en büyük 0.10 değerlerini alırken sola çarpık dağılan

büyük örneklemlerde en küçük 0.03, en büyük 0.05 değerlerini almıştır. Genel olarak örneklem büyüklüğünün artmasıyla birlikte standart hata değerlerinde azalma olduğu görülmektedir.

Tablo 5'te MHM'ye göre 30 madde için elde edilen maddelere ait SH değerleri; normal dağılan küçük örneklemlerde en küçük 0.05, en büyük 0.09 değerlerini alırken normal dağılan büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Sağa çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.05, en büyük 0.11 değerlerini alırken sağa çarpık dağılan büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Sola çarpık dağılan küçük örneklemlerde SH değerleri; en küçük 0.05, en büyük 0.10 değerlerini alırken büyük örneklemlerde en küçük 0.02, en büyük 0.05 değerlerini almıştır. Genel olarak örneklem büyüklüğünün artmasıyla birlikte standart hata değerlerinde azalma olduğu görülmektedir.

Kategori sayısı dikkate alındığında SH değerlerinde hem 10 maddelik hem de 30 maddelik testler için belli bir örüntü gözlenmemiştir. Diğer bir deyişle kategori sayısının değişmesinin SH değerleri üzerinde önemli bir etkisinin olmadığı sonucuna ulaşılmıştır. SH değerleri üzerinde en etkili faktörün örneklem büyüklüğü olduğu görülmektedir. Örneklem büyüklüğü arttıkça SH değerleri azalmaktadır.

Testlere Ait Model Veri Uyumu Değerleri ve Bu Değerler İçin Kestirilen Standart Hata Değerleri

Tablo 6'da testlere ait model veri uyumu değerleri ve bu değerler için kestirilen standart hata değerleri verilmiştir.

Tablo 6. Testlere Ait Model Veri Uyumu Değerleri ve Bu Değerler İçin Kestirilen Standart Hata Değerleri

		Normal Dağılan				Sağa Çarpık Dağılan				Sola Çarpık Dağılan			
		100 kişi		500 kişi		100 kişi		500 kişi		100 kişi		500 kişi	
		H	SH	H	SH	H	SH	H	SH	H	SH	H	SH
10 madde	3	0.33*	0.05	0.33*	0.02	0.32*	0.04	0.31*	0.02	0.35*	0.05	0.35*	0.02
	5	0.34*	0.05	0.34*	0.02	0.33*	0.04	0.43**	0.02	0.36*	0.05	0.36*	0.02
	7	0.33*	0.05	0.34*	0.02	0.33*	0.04	0.32*	0.02	0.36*	0.05	0.35*	0.02
30 madde	3	0.34*	0.03	0.35*	0.02	0.35*	0.03	0.34*	0.02	0.36*	0.04	0.36*	0.02
	5	0.36*	0.03	0.35*	0.02	0.35*	0.03	0.35*	0.02	0.36*	0.04	0.36*	0.02
	7	0.37*	0.04	0.36*	0.02	0.36*	0.03	0.35*	0.02	0.36*	0.03	0.36*	0.02

* düşük uyum, ** orta uyum

Tablo 6'da verilen testlere ait model veri uyumu değerleri Mokken'a (1971) göre belirlenen değerlendirme ölçütleriyle incelendiğinde, kısa testler için MHM'ye en iyi uyuma beş kategorili puanlanan maddeler için sağa çarpık dağılan 500 kişilik örneklemden ulaşılmıştır. Bu durumda testlerin MHM'ye uyumu orta düzeydedir. Diğer tüm durumlarda testlerin MHM'ye düşük düzeyde uyumlu oldukları görülmektedir. Kategori sayısı dikkate alındığında kısa testlerde genel olarak en iyi uyuma beş kategoriden ulaşılmıştır. Ancak burada H katsayısının sayısal değeri üzerinden yorum yapılmaktadır. Genel olarak testlerin MHM'ye uyum düzeylerinin Mokken'a (1971) göre belirlenen değerlendirme ölçütlerine göre düşük düzeyde oldukları unutulmamalıdır.

Benzer inceleme 30 maddelik uzun testler için yapıldığında tüm test koşullarının MHM'ye uyum düzeyinin düşük düzeyde olduğu görülmektedir. Kategori sayısı dikkate alındığında uzun testlerde genel olarak en iyi uyuma yedi kategoriden ulaşılmıştır. Ancak burada H katsayısının sayısal değeri üzerinden yorum yapılmaktadır. Genel olarak testlerin MHM'ye uyum düzeylerinin Mokken'a (1971) göre belirlenen değerlendirme ölçütlerine göre düşük düzeyde oldukları unutulmamalıdır.

Tablo 6'da verilen SH değerleri incelendiğinde hem kısa hem de uzun testlerde örneklem büyüklüğünün artmasıyla birlikte bu değerlerin azaldığı görülmektedir. Genel olarak kategori sayısı SH değerleri üzerinde etkili değildir.

Testler İçin Kestirilen Güvenirlik Değerleri (α, λ ve LCRC)

Tablo 7'de testler için kestirilen α, λ ve LCRC güvenirlik değerleri verilmiştir.

Tablo 7. Testler İçin Kestirilen Güvenirlik Değerleri

Dağılım Şekli	Örneklem Büyüklüğü	Güvenirlik Katsayısı	10 madde			30 madde		
			3	5	7	3	5	7
Normal Dağılan	100	α	0.74	0.74	0.73	0.91	0.91	0.92
		λ	0.75	0.76	0.75	0.92	0.92	0.93
		LCRC	0.80	0.79	0.78	0.94	0.93	0.93
	500	α	0.75	0.75	0.75	0.91	0.91	0.92
		λ	0.75	0.75	0.76	0.91	0.92	0.92
		LCRC	0.79	0.78	0.78	0.92	0.92	0.93
Sağa Çarpık Dağılan	100	α	0.73	0.73	0.76	0.91	0.91	0.92
		λ	0.74	0.75	0.77	0.91	0.92	0.92
		LCRC	0.80	0.79	0.79	0.93	0.93	0.93
	500	α	0.73	0.79	0.74	0.91	0.91	0.92
		λ	0.74	0.80	0.75	0.91	0.92	0.92
		LCRC	0.77	0.82	0.77	0.92	0.92	0.92
Sola Çarpık Dağılan	100	α	0.76	0.76	0.77	0.91	0.92	0.92
		λ	0.77	0.77	0.78	0.91	0.92	0.92
		LCRC	0.82	0.81	0.80	0.93	0.93	0.93
	500	α	0.76	0.76	0.77	0.91	0.92	0.92
		λ	0.76	0.77	0.78	0.91	0.92	0.92
		LCRC	0.79	0.79	0.79	0.92	0.93	0.93

Tablo 7'de verilen testlere ait güvenirlik kestirimleri incelendiğinde örneklemlerin tüm dağılım şekillerinde 30 maddeden oluşan uzun testlerden elde edilen güvenirlik katsayılarının, 10 maddeden oluşan kısa testlerden elde edilen güvenirlik katsayılarına göre daha yüksek olduğu görülmektedir. Kategori sayısının değişmesiyle birlikte güvenirlik kestirimlerinin birbirlerinden çok farklı değerler almadıkları görülmektedir. Diğer bir deyişle kategori sayısının MHM'ye göre yapılan güvenirlik kestirimlerinde etkili bir faktör olduğu söylenemez. Belirlenen test koşullarında güvenilirlik kestirimi üzerinde en etkili faktörün madde sayısı olduğu görülmektedir. Madde sayısının artmasıyla birlikte tüm test koşullarında güvenirlik katsayıları artmıştır. Ayrıca Tablo 7'de görüldüğü gibi genel olarak α güvenirlik katsayısı, λ ve LCRC güvenirlik katsayılarının altında değerler vermiştir.

SONUÇLAR ve TARTIŞMA

Bu araştırmada simülatif veriler kullanılarak kategori sayısının maddelerin psikometrik özellikleri üzerindeki etkisinin POMTK modellerinden MHM ile kestirimlerinin incelenmesi amaçlanmıştır. Araştırmanın amacı doğrultusunda iki farklı büyüklükteki çeşitli dağılım özelliklerine sahip (normal dağılan, sağa çarpık dağılan ve sola çarpık dağılan) örneklemler için iki farklı test uzunluğunda (10 madde (kısa) ve 30 madde (uzun)), üç farklı sayıda kategoriye (üç, beş ve yedi) sahip maddeler 100 replikasyon yapılarak simülatif olarak üretilmiştir.

Araştırmada oluşturulan test koşullarında örneklemlerin dağılım özelliği ve büyüklüğüne göre MHM'ye uyum göstermeyen maddelerin olduğu belirlenmiştir. Kısa testlerde en fazla uyumsuzluk yedi kategoride puanlanan maddeler için sağa çarpık dağılan küçük örneklemlerde gözlenmiştir. Burada dört maddenin MHM'ye uyum göstermediği belirlenmiştir. Uzun testlerde en fazla uyumsuzluk üç kategoride puanlanan maddeler için sağa çarpık dağılan büyük örneklemlerde gözlenmiştir. Burada altı maddenin MHM'ye uyum göstermediği belirlenmiştir. Bunların dışındaki diğer koşullarda maddelerin tamamına yakının MHM'ye uyum gösterdiği belirlenmiştir. Elde edilen bu sonuç araştırma kapsamında oluşturulan koşullarda küçük örneklemlere uygulanan testlerin MHM'ye uyumlu olduğunu göstermektedir. Araştırmanın bu bulgusu, literatürde belirtilen küçük

örneklemlere uygulanan kısa testlerin MHM'ye uyumlu olduğu bulgusuyla (Junker ve Sijtsma, 2001; Meijer, 2004; Molenaar, 2001) paralellik göstermektedir. Ayrıca araştırma kapsamında oluşturulan test koşullarında testlerden elde edilen H katsayıları incelendiğinde Mokken (1971) tarafından belirlenen değerlendirme ölçütlerine göre testlerin genelde MHM'ye düşük düzeyde uyumlu oldukları belirlenmiştir.

Araştırmada belirlenen test koşullarında maddelerin kategori sayısının değişmesiyle birlikte MHM'ye uyum düzeyinde belli bir örüntüye göre değişim gözlenmemiştir. Ancak genel olarak uyum düzeyi değişmemekle birlikte kategori sayısının artmasıyla H katsayılarının daha yüksek değerler aldığı gözlenmiştir. MHM'ye uyum göstergesi olarak tanımlanan H katsayılarının değerleri arttıkça daha ayırt edici maddelere ulaşılmaktadır (Sijtsma ve Molenaar, 2002). Bu bağlamda MHM'ye uyum düzeyinin artması, maddelerin ve testlerin psikometrik açıdan daha geçerli ölçme yapabildiğini göstermektedir. Yapılan bu araştırmada kategori sayısının artması testlerin MHM'ye uyumunu değiştirmemiştir. Diğer bir deyişle araştırma koşullarında kategori sayısının MHM'ye göre yapılan ölçeklemede geçerlik üzerinde etkisi olmamıştır. Araştırmanın bu bulgusu kategori sayısının artmasına bağlı olarak ölçme geçerliliğinin arttığını ifade eden KTK kapsamında yürütülen bazı araştırmalardan (Lozano, García-Cueto ve Muñiz, 2008; Preston ve Colman, 2000) farklı çıkmıştır. Ayrıca KTK kapsamında yürütülen ve kategori sayısının geçerliği etkilemediğini ifade eden araştırmalara (Erkuş, Sanlı, Bağlı ve Güven, 2000; Maydeu-Olivares ve diğerleri, 2009; Uyumaz ve Çokluk, 2016) paralel çıkmıştır. Yine araştırmanın bu bulgusu PMTK kapsamında yürütülen araştırmalardan İlhan ve Güler (2017) ve Lee ve Paek (2014) tarafından yürütülen araştırmanın bulgularına paralel çıkarken Fabiola ve diğerlerinin (2012) ve Maydeu-Olivares ve diğerlerinin (2009) bulgularından farklılaşmıştır.

Araştırmada maddeler ve testler için kestirilen SH değerleri incelendiğinde, kategori sayısının bu değerler üzerinde etkili olmadığı belirlenmiştir. SH değerleri için yapılan kestirimler örneklem büyüklüğünün artmasıyla birlikte azalmıştır. Diğer bir deyişle örneklem büyüklüğü arttıkça H kestirimleri için yapılan hata değerleri azalmıştır. Araştırmanın bu bulgusu literatüre paraleldir (Smits, Timmerman ve Meijer, 2012; Şengül Avşar ve Tavşancıl, 2017; Koğar, 2015).

Araştırmada madde sayısına bağlı olarak tüm güvenirlik değerlerinin arttığı görülmüştür. Madde sayısının artmasıyla birlikte güvenirliğin de arttığı (Crocker ve Algina,1986) bilinmektedir. Ayrıca araştırmanın pek çok koşulunda kategori sayısının artmasıyla birlikte güvenirlik değerlerinde küçük artışlar gözlenmiştir. Araştırmanın bu bulgusu KTK kapsamında yürütülen araştırma bulgularına paraleldir (Erkuş, Sanlı, Bağlı ve Güven, 2000; Leung, 2011; Lozano, García-Cueto ve Muñiz, 2008; Maydeu-Olivares ve diğerleri, 2009; Preston ve Colman, 2000, Uyumaz ve Çokluk, 2016; Weng, 2004). PTMK kapsamında yürütülen araştırmalarda da (Maydeu-Olivares ve diğerleri, 2009; Lee ve Paek, 2014; Pozehl, 1990; Wang, 2004; Zenisky, Hambleton ve Sireci, 2002; Zhang, 2010) benzer sonuçlara ulaşılmıştır.

Çok kategorili puanlanan maddelerden oluşan testlerin, MHM'ye göre kestirilen bir başka güvenirlik değeri olan LCRC'nin de yüksek kestirildiği görülmüştür. Bu bulgu Rivas, Bersabé ve Berrocal'ın (2005) yaptıkları araştırmada belirtilen MHM'ye göre yapılan ölçekleme ile yüksek güvenirlikte testlere ulaşılacağı bulgusunu desteklemektedir. Araştırmada en yüksek güvenirlik kestirimlerine küçük farklarla da olsa hem kısa hem de uzun testlerde genellikle beş ve yedi kategorili puanlanan maddelerden ulaşılmıştır. Araştırmanın bu sonucu, araştırmalarında aynı kategori sayılarını seçen İlhan ve Güler'in (2017) araştırma bulgusuna paraleldir.

Araştırmadan elde edilen sonuçlar özetlendiğinde kategori sayısının MHM'ye göre yapılan kestirimler üzerinde çok büyük etkilerinin olmadığı sonucuna ulaşılmıştır. Genellikle kategori sayısının artmasıyla birlikte geçerlik katsayısı gibi yorumlanabilen H katsayılarının MHM'ye uyum düzeylerini değiştirmediği ancak bazı durumlarda uyum düzeyi aynı kalmakla birlikte daha yüksek değerler aldığı görülmüştür. Kategori sayısının güvenirlik kestirimleri için seçilen α, λ ve LCRC katsayıları üzerinde etkili bir faktör olmadığı belirlenmiştir. Madde sayısı arttıkça α, λ ve LCRC katsayıları yüksek değerler almıştır. Ayrıca kategori sayısı, maddeler ve testler için kestirilen SH değerleri üzerinde etkili bulunmamıştır.

Araştırmadan elde edilen sonuçlardan POMTK modellerinden MHM'nin kısa testlerde ve küçük örneklemlerde uyumlu olduğu görülmüştür. Bu nedenle küçük örneklemlere ulaşabilen araştırmacılar için verilerini MHM'ye göre ölçeklemeleri önerilebilir. Araştırma sonuçlarından özellikle çok kategorili puanlanan maddelerden oluşan ve MHM'ye göre ölçekleme yaparak ölçek geliştirmek isteyen araştırmacıların beş kategorili puanlanan maddeleri kullanmaları önerilebilir. Ancak bu önerilerde araştırmanın sınırlılıkları göz önünde bulundurulmalıdır.

Araştırma kapsamında getirilen önerilere ek olarak simülatif bir çalışma olarak yürütülen bu araştırmada belirlenen test koşulları dışında, farklı madde ve kategori sayılarına sahip testler, değişen örneklem büyüklükleri ve değişen çarpıklık durumları gibi farklı test koşulları oluşturularak yeni çalışmalar yapılabilir. Gerçek veri setleriyle de benzer bir çalışma yapılarak bu araştırmanın sonuçlarıyla karşılaştırılabilir.

KAYNAKÇA

Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of nonnormal distributions and small sample size* (Master's Thesis). Available from ProQuest Dissertations and Theses database. (UMI No. MR90146)

Bulmer, M. G. (1979). *Principles of statistics.* New York: Dover Publications.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, *17*(4), 335-350. doi:10.1177/01466216930170040

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando: Harcourt Brace Jovanovich Inc.

DeMars, C. E. (2002, April). *Recovery of graded response and partial credit parameters in Multilog and Parscale.* Paper presented at the annual meeting of American Educational Research Association, Chicago.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*(3), 224- 247. doi:10.1177/0146621607302479

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

Erkuş, A., Sanlı, N., Bağlı, M. ve Güven, K. (2000). Öğretmenliğe ilişkin tutum ölçeği geliştirilmesi. *Eğitim ve Bilim, 25*(116), 27-33. http://egitimvebilim.ted.org.tr/index.php/EB/article/view/5276/1439 adresinden erişildi.

Fabiola, G., Iwin, L., Jennifer, L., & Zaira, V. (2012). The effect of the number of answer choices on the psychometric properties of stress measurement in an instrument applied to children. *Evaluar*, *12*, 43-59. Retrieved from https://revistas.unc.edu.ar/index.php/revaluar/article/viewFile/4694/4488

Galindo-Garre, F., Hendriks, S. A., Volicer, L., Smalbrugge, M., Hertogh, C. M., & van der Steen, J. T. (2014). The bedford alzheimer nursing-severity scale to assess dementia severity in advanced dementia: A nonparametric item response analysis and a study of its psychometric characteristics. *American Journal of Alzheimer's Disease and Other Dementias, 29*(1), 84-90. doi:10.1177/1533317513506777

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement, 12*, 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*(4), 679-693.

İlhan, M., & Güler, N. (2017). The number of response categories and the reverse directional item problem in likert-type scales: A study with the rasch model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 321-343. doi:10.21031/epod.321057

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 109, 1-10. doi:10.3389/fpsyg.2016.00109

Junker, B., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, *25*(3), 211- 220. doi:10.1177/01466210122032028

Koğar H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir monte carlo çalışması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*, 142-157. doi:10.21031/epod.02072

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item characteristic curves for binary items. *Applied Psychological Measurement, 31*(2), 121–134. doi:10.1177/0146621606290248

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment, 32*(7), 663-673. doi:10.1177/0734282914522200

Lee, Y. S., Wollack, J. A., & Douglas, J. (2009). On the use of nonparametric item characteristic curve estimation techniques for checking parametric model fit. *Educational and Psychological Measurement, 69*(2), 181–197. doi:10.1177/0013164408322026

Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of Social Service Research, 37*(4), 412-421. doi:10.1080/01488376.2011.580697

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement, 51*(1), 1–17. doi:10.1111/jedm.12031

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*: *European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73-79. doi:10.1027/1614-2241.4.2.73

Maydeu-Olivares, A., Kramp, U., García-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods, 41*(2), 295-308. doi:10.3758/BRM.41.2.295

Meijer, R. R. (2004, March). *Investigating the quality of items in cat using nonparametric IRT* (Report No. 04-05). Law School Admission Council Computerized Testing Report. A Publication of the Law School Admission Council.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*(3), 354-368. doi:10.1037/1082-989X.9.3.354

Mokken, R. J. (1971). *A theory and procedure of scale analysis: With applications in political research*. The Hague: Mouton.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-368). New York: Springer-Verlag.

Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 295-299. doi:10.1177/01466210122032091

Ostini, R., & Nering, M. L. (2006*). Polytomous item response theory models*. Thousand Oaks, CA: Sage

Patsula, N. L., & Gessaroli, E. M. (April, 1995). *A comparison of item parameter estimates and iccs produced with testgraf and bilog under different test lengths and sample sizes.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Pozehl, J. B. (1990). *Application of item response theory to criterion-referenced measurement: An investigation of the effects of model choice, sample size, and test length on reliability and estimation accuracy* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9030146)

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15. doi:10.1016/S0001-6918(99)00050-5

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Rivas, T., Bersabé, R., & Berrocal, C. (2005). Application of double monotonicity model to polytomous items: Scalability of the beck depression items on subjects with eating disorders. *European Journal of Psychological Assessment, 21*(1), 1-10. doi:10.1027//1015-5759.21.1.1

Sachs, J., Law, Y. K., & Chan, C. K. K. (2003). A nonparametric item analysis of a selected item subset of the learning process. *British Journal of Educational Psychology, 73*(3), 395–423. doi:10.1348/000709903322275902

Sijtsma, K., & Molenaar, W. I. (2002). *Introduction to nonparametric item response theory*. USA: Sage Publications.

Sijtsma, K., Debets, P., & Molenaar, W. I. (1990). *Mokken scale analysis for polychotomous items: Theory, a computer program and an empirical application.* Netherlands: Quality and Quantity, Kluwer academic publishers.

Smits, I. A. M., Timmerman, M. E., & Meijer, R. R. (2012). Exploratory mokken scale analysis as a dimensionality assessment tool: Why scalability does not imply unidimensionality. *Applied Psychological Measurement, 36*(6), 516-539. doi:10.1177/0146621612451050

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two parameter logistic response model: An evaluation of Multilog. *Applied Psychological Measurement, 16*(l), 1-16. doi:10.1177/014662169201600101

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*(4), 331-352. doi:10.1111/j.1745-3984.2003.tb01150.x

Štochl, J. (2007). Nonparametric extension of item response theory models and its usefulness for assessment of dimensionality of motor tests. *Acta Universitatis Carolinae*, *42*(1), 75-94.

Sueiro, M. J., & Abad, F. J. (2011). Assessing goodness of fit in item response theory with nonparametric models: A comparison of posterior probabilities and kernel smoothing approach. *Educational and Psychological Measurement, 71*(5), 834-848. doi:10.1177/0013164410393238

Syu, J. J. (2013). *Applying person fit-in faking detection-the simulation and practice of non parametric item response theory*. (Doctoral Dissertation, National Chengchi University). Retrieved from http://nccur.lib.nccu.edu.tw/bitstream/140.119/58646/1/251501.pdf

Şengül Avşar, A., & Tavşancıl, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory & Practice*, *17*(2). doi:10.12738/estp.2017.2.0246

Tendeiro, J. N., & Meijer, R. R. (2013). The probability of exceedance as a nonparametric person fit statistic for tests of moderate length. *Applied Psychological Measurement, 37*(8), 653–665. doi:10.1177/0146621613499066

Uyumaz, G., & Çokluk, Ö. (2016). An investigation of item order and rating differences in likert-type scales in terms of psychometric properties and attitudes of respondents. *Journal of Theoretical Educational Science, 9*(3), 400-425. doi:10.5578/keg.10011

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1-19.

van der Ark, L. A. (2015). Package 'mokken'. Retrieved from http://cran.rproject.org/web/packages/mokken/mokken.pdf

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*(5), 380-392. doi:10.1177/0146621610392911

van Onna, M. J. H. (2004). Estimates of the sampling distribution of scalability coefficient h. *Applied Psychological Measurement, 28*(6), 427-449. doi:10.1177/0146621604268735

Wang, W. C. (2004). Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Educational and Psychological Measurement*, *64*(6), 937-955. doi:10.1177/0013164404268671

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956-972. doi:10.1177/0013164404268674

Young, M. A., Blodgett, C., & Reardon, A. (2003). Measuring seasonality: Psychometric properties of the seasonal pattern assessment questionnaire and the inventory for seasonal variation. *Psychiatry Research, 117*(1), 75-83. doi: 10.1016/S0165-1781(02)00299-8

Zhang, O. (2010). *Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations* (Master's Thesis, University of Florida). Retrieved from http://ufdc.ufl.edu/UFE0042638/00001

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*(4), 291 -309. doi:10.1111/j.1745-3984.2002.tb01144.x

EXTENDED ABSTRACT

Introduction

Measuring the affective properties of individuals takes an important place in education and psychology. It is seen that the measurement tools used to measure these properties are mostly composed of items that have response categories in which graded or polytomous response are presented, rather than items that are scored binary.

The identification of the psychometric properties of measurement tools is very important in terms of the accuracy and suitability of the decisions made according to these tools. Psychometric properties are determined by various test theories such as Classical Test Theory (CTT) and Item Response

Theory (IRT). These theories are often used in educational and psychological research. With the contributions of recent studies conducted in the literature, IRT has been classified as Parametric Item Response Theory (PIRT) and Non-parametric Item Response Theory (NIRT) (Sijtsma and Molenaar, 2002).

NIRT models are models that provide convenience in applications for short tests and small samples. Upon analysing the literature, it is seen that NIRT models can be classified as Mokken model and nonparametric regression models (Şengül Avşar and Tavşancıl, 2017). Mokken model is divided into sub-models as Monotone Homogeneity Model (MHM) and Double Monotonicity Model (DMM) (Sijtsma and Molenaar, 2002).

MHM, which is one of the NIRT models, is a statistical measurement model that requires fewer assumptions compared to PIRT models. MHM requires unidimensionality, local independence, and monotonicity assumptions (Sijtsma and Molenaar, 2002). It is seen clearly that MHM requires similar assumptions as unidimensional PIRT models. However, when the PIRT and NIRT models are compared, it is known that the essential difference between the two models depends on the item characteristic curves (ICC) (Şengül Avşar and Tavşancıl, 2017). While ICCs are monotonically and logistically estimated in PIRT models, these curves cannot be estimated in NIRT models logistically, even though they are monotone (Sijtsma and Molenaar, 2002).

Parameter estimations according to MHM are made with "scalability coefficient (H)". The H coefficient, which is developed by Loevinger (1947,1948) for the binary scored items has been regulated by Mokken (1971) to identify the relationship of a single item with other items in the set (H_i) and the entire set of items (H) for the item pairs (i, j item pairs) (H_{ij}) that are in a set in MHM (Mokken, 1997). Aside from its being interpreted as the nonparametric equivalent of the coefficient "a" (item discrimination index) in the logistic models with two or three parameters, the H coefficient also means scalability index which is employed in determining whether the measurement tools are scaled according to MHM (Meijer, 2004; Mokken, 1997; van Onna, 2004). Reliability estimations in MHM are made with Cronbach α, Guttman lambda 2 (λ), and latent class reliability coefficient (LCRC).

It is observed in the literature review that several studies in which the psychometric properties of the scales are examined according to the NIRT models are carried out in Turkey and abroad (Galindo Garre et al., 2014; Sachs, Law, and Chan, 2003; Koğar, 2015; Şengül, Avşar, and Tavşancıl, 2017; Young, Blodgett and Reardon, 2003). Especially in the studies on the polytomous items, researches are carried out on both simulated and real data sets according to various factors such as sample size, the number of items and distribution of sample.

Another factor affecting the psychometric properties of measurement tools, which are composed of polytomous items and used in the measurement of affective properties, is the number of categories (Leung, 2011; Lozano, García-Cueto, and Muñiz, 2008; Preston and Colman, 2000; Weng, 2004). In this context, it is seen that various studies investigating the effect of category number on psychometric properties within the scope of CTT and PIRT have been made.

It is seen in the literature that MHM is a useful model in determining the psychometric properties of measurement tools and the number of categories is an important factor in analysing the psychometric properties of measurement tools consisting of polytomous items. It has been found necessary to determine the psychometric properties of the measurement tools with different category numbers via the simulated data sets with the MHM from the NIRT models.

Method

This research is a basic research which aims to determine the effects of the number of categories on the psychometric properties of tests consisting of polytomous items via simulated data.

The study consists of 36 different test conditions; two different sample size (100 and 500), three different sample distribution shapes (normal distribution, positively skewed distribution and

negatively skewed distribution), two different test lengths (10 items and 30 items), three number of categories (three, five and seven). The conditions and the selection of the parameters required for the generation of data in the study are determined by examining similar studies that are conducted in the literature. The data sets are produced in the conditions indicated in the research by replicating 100 times.

R Studio 3.4.0 software is employed to generate the simulated data. R Studio 3.4.0 software was used for making analyses of the data according to the MHM and for the LCRC reliability coefficient calculations, and the Mokken package which was developed by van der Ark (2007) was used in the R Studio software.

Results and Discussion

When the results gathered from the study are summarized, it was observed that there was not any specific pattern of item fit to MHM with changing the number of categories of polytomous items. It was found that tests have weak fit to MHM under test conditions in the research. In general, it was seen that the number of categories has no effect on reliability estimates for both short and long tests. Reliability coefficients α, λ, and LCRC are estimated as having higher values for long tests than short tests. Nevertheless, it should be pointed out that the reliability estimations are generally high, and the differences emerged depending on the number of categories are not immense. Furthermore, the number of categories was not found to be influential in standard error values estimated for the tests. In addition to the fact that findings obtained from the study are obtained from estimations made with MHM, which is a NIRT model, the findings are also parallel to some findings in the literature obtained from the analyses conducted with PIRT and CTT.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(1);64-84

Öğrenci, Öğretmen ve Öğretimsel Nitelikler Açısından TIMSS-2015'e Dayalı Olarak Öğrencilerin Sınıflandırılması

Classification of Students In Terms of Student's, Teacher's and Instructional Qualifications Based on TIMSS-2015

Emine ÖNEN*

Öz

Bu araştırmanın amacı, TIMSS-2015 uygulamasına dayalı olarak, matematik başarısını etkileyebileceğini düşünülen öğrenci ve öğretmen nitelikleri ile öğretimsel nitelikler açısından dördüncü ve sekizinci sınıf öğrencilerini sınıflandırarak öğrenci profilleri oluşturmaktır. Yapılan kümeleme analizi sonucunda dördüncü sınıf düzeyinde üç kümenin ortaya çıktığı, sekizinci sınıf düzeyinde ise iki kümenin ortaya çıktığı gözlenmiştir. Dördüncü sınıf düzeyi için bu kümelerin oluşmasında en etkili olan özellikler, matematik başarısı, öğrenci ve öğretmen nitelikleridir. Öğretmen beyanına dayalı belirlenen öğretimsel niteliklerin sınıflama işleminde çok az önemli olduğu bulunmuştur. Sekizinci sınıf düzeyi için ise öğrenci niteliklerinin bu sınıflamada en etkili faktörler olduğu ancak matematik başarısının ve öğretmen nitelikleri ile öğretimsel niteliklerin, öğrencilerin sınıflanmasında düşük düzeyde önemli olduğu görülmüştür. Tanımlanan öğrenci profilleri, öğrenci nitelikleri açısından her iki sınıf düzeyinde de benzerlik gösterirken; öğretmen nitelikleri ve öğretimsel nitelikler açısından dördüncü ve sekizinci sınıf düzeyi için farklılaşmaktadır. Her iki sınıf düzeyinde de matematikte en başarılı öğrencilerin matematikte kendine güven düzeyi çok düşük, matematik öğrenmeyi seven, matematik dersindeki öğretimin ilgi çekici olduğunu düşünen, okula aidiyet hissi yüksek ve akran baskısına çok az maruz kalan öğrenciler olduğu gözlenmiştir. Matematikte başarı düzeyi düşük öğrencilerin ise matematikte kendine çok güvenen, matematik öğrenmeyi sevmeyen, matematik dersindeki öğretimi ilgi çekici olmadığını düşünen, okula aidiyet hissi düşük ve akran baskısına maruz kalan öğrenciler olduğu görülmektedir.

Anahtar Kelimeler: Geniş ölçekli testler, Öğrenci çıktıları, Matematik başarısı, Öğretmen kalitesi, Öğretimsel kalite

Abstract

The aim of this study is creating students' profiles by clustering them based on the student and teacher attributions and instructioanal qualifiactions that could affect their mathematics achievement. As a result of cluster analysis, it has been observed that, three classes at the forth-grade level have emerged and two classes at the eight-grade level have emerged. It has been found that the instructional qualifications determined based on the teachers' reports have little importance at this classification. For the eight grade level, it is seen that student attributions as the most important factors at this classification but mathematics achievement, teacher attributions and instructional qualifications are seen as having little effects on this classification at the both grade levels, it has been observed that the most successful students in mathhematics are the ones; whose confidence level are too low in math, who like learning math, who think in math lessons engaging teaching is put into practice, whose sense of school belonging are high and who slightly are exposed to peer pressure. On the other hand, it is seen that students with low-level math achievement are the ones; who have confidence in math so much, who do not like learning math, who think instruction is not interesting in math lessons, whose sense of school belonging are so low and who are highly exposed to peer pressure.

* Dr. Öğretim Üyesi,; Gazi Üniversitesi, Gazi Eğitim Fakültesi, Eğitim Bilimleri Bölümü, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı, Ankara, Türkiye. E-posta: emine_onen@yahoo.com, ORCID ID: orcid.org/0000-0002-0398-3191

DOI: 10.21031/epod.364319
Geliş Tarihi: 12.10.2017
Kabul Tarihi: 15.02.2018

Keywords: Large-scale tests, student outcomes, mathematics achievement, teacher quality, instructional quality

GİRİŞ

Geniş ölçekli testler, son 30 yıldır dünya genelinde ülkeler tarafından eğitim sistemleri hakkında bilgi edinmek ve eğitim sistemlerinin gelişimini/değişimini izlemek üzere uygulanmaktadır. Geniş ölçekli testler eğitim politikalarına ilişkin karar verme, hesap verebilirlik ve eğitimsel planlama açısından önemli bir rol oynamaktadır (Ercikan, Simon, Oliveri, 2013). Bu tür testler çeşitli amaçlar için uygulanmasına rağmen, en yaygın şekilde eğitim sisteminin çıktılarını rapor etmek ve buna dayalı olarak eğitim sisteminin kalitesini sağlamak amacıyla uygulandığı görülmektedir. Geniş ölçekli testler, eğitim sistemlerinin zayıf ve güçlü yönleri hakkında ayrıntılı bilgi vermekte ve sistemin zaman içerisindeki değişiminin-gelişiminin izlenmesini sağlamaktadır. Bu doğrultuda sistemde eksiklik ve sorunları gidermeye ve nihayetinde de sistemin kalitesini arttırmaya yönelik müdahaleler yapılmaktadır (Tobin, Lietz, Nugroho, Vivekanandan ve Nyamkhuu, 2015).

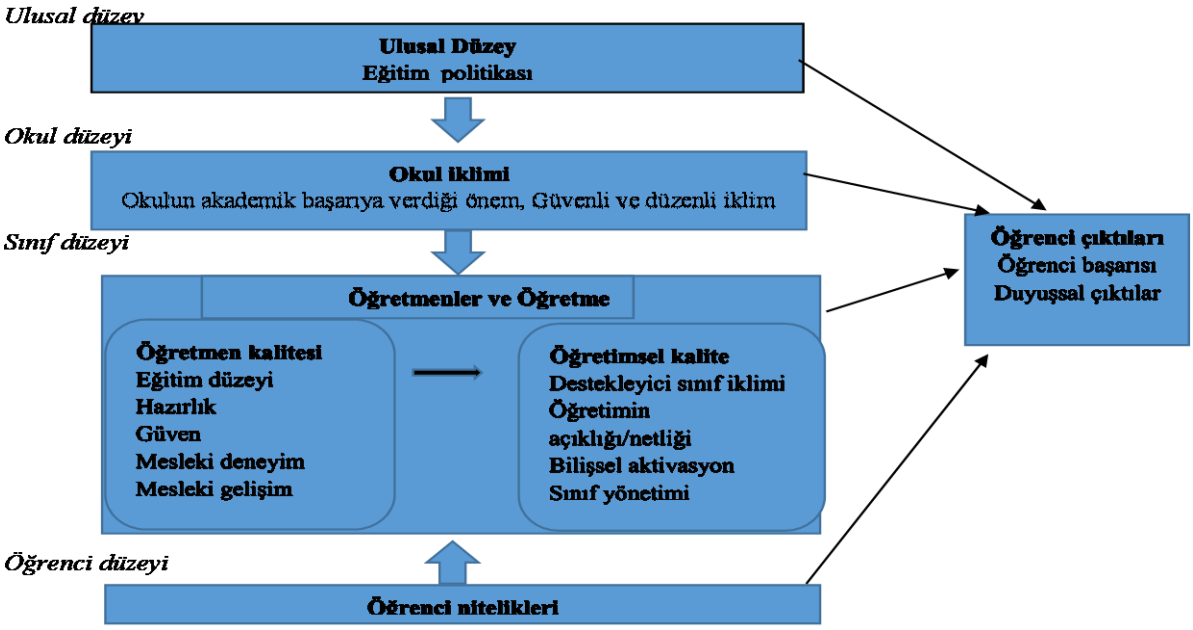
Geniş ölçekli testler arasında, uluslararası düzeyde uygulanan PISA (Programme for International Student Assessment: Uluslararası Öğrenci Değerlendirme Programı), PIRLS (Progress in International Reading Literacy Study: Uluslararası Okuma Becerilerinde Gelişim Çalışması), TALIS (Teaching and Learning International Survey: Uluslararası Öğretme ve Öğrenme Araştırması) ve TIMSS yer almaktadır (Trends in Mathematics and Science Study: Uluslararası Matematik ve Fen Eğilimleri Araştırması). Bu araştırma TIMSS-2015 verilerine dayalı olarak gerçekleştirildiğinden burada yalnızca TIMSS hakkında ayrıntılı bilgi verilmiştir. TIMSS, öğrencilerin matematik ve fen alanlarında kazandıkları bilgi ve becerilerin değerlendirilmesi amacıyla Uluslararası Eğitimsel Başarıyı Değerlendirme Kuruluşu (International Association for the Evaluation of Educational Achievement: IEA) tarafından geliştirilmekte ve uygulanmaktadır. İlk olarak 1995 yılında uygulanan TIMSS, ardından dörder yıllık dönemler halinde katılımcı ülkelerdeki dördüncü ve sekizinci sınıf düzeyindeki öğrencilere uygulanmıştır (Martin, Mullis ve Hooper, 2016). Türkiye ise ilk olarak, 1999 yılında uygulanmış olan TIMSS-R'a (The Third International Mathematics and Science Study–Repeat: Üçüncü Uluslararası Matematik ve Fen Bilimleri Çalışması) sadece sekizinci sınıf düzeyinde katılmıştır. Türkiye, 2003 yılındaki TIMSS uygulamasına katılmamış, 2007 yılındaki uygulamaya sadece sekizinci sınıf düzeyinde katılmış, 2011 ve 2015 yıllarındaki uygulamalara ise hem dördüncü sınıf hem de sekizinci sınıf düzeyinde katılmıştır **(**Yıldırım, Özgürlük, Parlak, Gönen ve Polat, 2016**)**.

Öğrencilerin matematik ve fen alanlarındaki bilgi ve beceri düzeyleri, öğretim programları, öğrenci nitelikleri, öğretmen ve okul niteliklerine ilişkin bilgiler TIMSS çalışması ile sağlamaktadır. Elde edilen bu bilgilere dayalı olarak, matematik ve fen eğitiminin kalitesini arttırmak üzere ulusal ve uluslararası düzeyde karşılaştırmalı olarak öğretim programlarının ve yöntemlerinin değerlendirilmesi mümkün olmaktadır (Martin ve diğerleri., 2016). TIMSS, dünyadaki en büyük ve en kapsamlı uluslararası öğrenci başarılarını değerlendirme çalışmasıdır ve geniş ölçekli testler içerisinde sadece TIMSS öğrenci, sınıf ve okul düzeyinde bilgi sağlamaktadır. Dolayısıyla TIMSS, araştırmacılara zaman içerisinde ve ülkeler arası karşılaştırmalı olarak, öğrenci ile öğretmen nitelikleri, öğretimsel nitelikler ve öğrenme çıktıları arasındaki ilişkileri çalışmak için fırsatlar sunmaktadır (Nilsen, Gustaffson ve Blömeke, 2016). TIMSS, bu yönüyle araştırmacılara dünyayı bir eğitim laboratuvarı olarak kullanma imkânı sağlamaktadır. Dolayısıyla bu bağlamda öğretmen kalitesine ve öğretimsel kaliteye ilişkin uluslararası bir anlayışın gelişmesine katkı sağlayarak bunların öğrenci öğrenme çıktıları için önemini ortaya koymaktadır.

TIMSS verilerine dayalı olarak ulusal ve uluslararası düzeyde gerçekleştirilen çalışmaların (Akyüz, 2006; Berberoğlu, Çelebi, Özdemir, Uysal ve Yayan, 2003; Blömeke, Olsen ve Suhl, 2016; Bos ve Kuiper, 1999; Gustafsson ve Nilsen, 2016; House ve Telese, 2008; Rutkowski ve Rutkowski, 2016) sonuçları da öğrenci öğrenme çıktılarının, öğretmene ve öğretim sürecine ilişkin faktörlerle ilişkilerini ortaya koymaktadır. İlgili alanyazın incelendiğinde de öğrenci çıktıları ile öğretmen

nitelikleri, öğretimsel nitelikler ve okul nitelikleri arasındaki ilişkileri açıklamaya yönelik kuramsal modellerin geliştirildiği görülmektedir.

Creemers ve Kyriakides'in (2006) önerdiği "Eğitimsel Etkililiğin Dinamik Modeli", kapsamlı ve uluslararası düzeyde iyi bilinen modellerden bir tanesidir. Bu model öğrenci çıktılarını etkileyebilecek olası faktörleri; (a) ulusal düzey, (b) okul düzeyi, (c) sınıf düzeyi ve (d) öğrenci düzeyi olmak üzere dört düzeyde ele almaktadır (Akt: Creemers ve Kyriakides, 2009). Nilsen ve diğerleri (2016) ise bu modele dayalı olarak, öğrenci çıktılarının belirleyicilerine ilişkin bir kavramsal çerçeve oluşturmuşlar ve yayınladıkları eserde bu kavramsal çerçeveyi esas alan ve TIMSS-2011 uygulamasına dayalı olarak gerçekleştirilen çalışmaları sunmuşlardır. Öğrenci çıktılarının belirleyicilerini ulusal düzeyde, okul düzeyinde, sınıf düzeyinde ve öğrenci düzeyinde ele alarak, eğitim sistemlerinin kompleks yapısını ortaya koyan bu kavramsal çerçeve aşağıda Şekil 1'de sunulmaktadır.



Şekil 1. Öğrenci Çıktılarının Belirleyicilerine İlişkin Kavramsal Çerçeve

Bu kavramsal çerçeve, ulusal düzey, okul düzeyi, sınıf düzeyi ve öğrenci düzeyi arasındaki doğrudan ve dolaylı ilişkilere odaklanmaktadır. Ulusal düzeyde kültürel faktörler, eğitsel değerler, eğitim politikaları ve okul sistemlerini de içerecek şekilde eğitim sistemlerindeki farklılıklar ele alınmaktadır. Bu farklılıkların daha alt düzeyde okulları, sınıfları ve öğrencileri etkilediği varsayılmaktadır. Okul düzeyinde, "okulun akademik başarıya verdiği önem" ve "güvenli ve düzenli okul iklimi algısı" faktörleri dikkate alınmıştır. Sınıf düzeyi, özellikle öğrenci öğrenme çıktıları için önemli olduğu düşünülen iki boyutu içermektedir: (a) Öğretmen kalitesi ve (b) Öğretimsel kalite (Gustafsson ve Nilsen, 2016). Öğretmen, okul ortamlarında öğrenmeyi sağlamak üzere öğrencinin çevresi ile etkileşimini planlayan ve düzenleyen kişi konumundadır. Dolayısıyla öğretmenin niteliklerinin öğrenci çıktıları üzerindeki etkisinin derecesi eğitimsel sistemlere göre değişmekle birlikte, öğretmenin nitelikleri öğrenci başarısı açısından oldukça önemlidir. Öğretmen kalitesi, öğretmene ilişkin çeşitli niteliklere işaret etmektedir: Öğretmenin eğitimsel geçmişi bağlamında öğrenim düzeyi ile öğretmenlik mesleğindeki deneyimi, mesleki gelişim etkinliklerine katılımı ile öğretmenin mesleki boyutta özyeterlik inancı ve kişilik özellikleri (Blömeke ve diğerleri., 2016). Bu konudaki araştırmaların (Akyüz, 2006; Blömeke ve diğerleri., 2016; Buluç, 2014; Butakor, 2016; Gustafsson ve Nilsen, 2016; Hong, 2012) sonuçları da bu tür öğretmen niteliklerinin, öğrenci

çıktılarıyla ilişkilerini ortaya koymaktadır. Öğretmene ilişkin bu niteliklerin aynı zamanda sınıf içi öğretimsel süreçler için de kaynak olduğu vurgulanmaktadır.

Öğretimsel kalite ise, alanyazında operasyonel olarak farklı şekillerde tanımlanmaktadır. TIMSS'te ise öğretimsel kalitenin bilişsel aktivasyon (cognitive activation), öğretimin açıklığı/netliği ve destekleyici (sınıf) iklim yönleriyle ele alınıp ölçüldüğü belirtilmektedir. Burada esas alınan kavramsal çerçevede öğretimsel kalite, bu üç boyutuyla ele alınmıştır (Gustafsson ve Nilsen, 2016). TIMSS'te öğretimsel kalite hem öğretmen hem de öğrenci algısına dayalı olarak ölçülmektedir. Öğretmen anketinde yer alan ve öğretmenlere, belirtilen etkinlikleri sınıf içerisinde hangi sıklıkta gerçekleştirdiklerinin sorulduğu maddeler yoluyla öğretimsel kalitenin bu üç yönüne ilişkin bilgi edinilmektedir. Öğrenci anketinde yer alan ve öğrencilere o alandaki (matematik/fen) öğretimi ne düzeyde ilgi çekici olarak düşündüklerine yönelik maddeler aracılığıyla da öğrenci algısına dayalı olarak öğretimsel kaliteye ilişkin bilgi sağlanmaktadır. Öğretimsel niteliklerin öğrenci başarısı ile ilişkilerine yönelik kanıtlar, uluslararası ve ulusal düzeyde gerçekleştirilen çalışmalardan (Blömeke ve diğerleri., 2016; House ve Telese, 2008; Yayan, 2003) gelmektedir. Öğretmenin sınıf içi öğretim uygulamalarını etkileyen diğer bir faktör ise, öğretmenin diğer öğretmenlerle etkileşimidir. Bu doğrultuda bu araştırmada öğretmenlerle arası etkileşim düzeyi de öğretimsel kalitenin göstergelerinden biri olarak ele alınmıştır (Little, 1981).

Öğrenci düzeyinde ise, bilişsel ve duyuşsal boyutta öğrenci çıktılarını etkileyen cinsiyet, sosyo-ekonomik düzey gibi demografik özelliklerin yanısıra duyuşsal özellikler de dikkate alınmıştır. Bu konuda yapılan çalışmalara bakıldığında başarı ile en fazla ilişkili olan duyuşsal özellikler arasında o alanda öğrencinin kendine güvenmesi (Akyüz, 2014; Butakor, 2016; Stemler, 2001; Yoshino, 2012), o alanı öğrenmeyi sevmesi (Akyüz, 2014; Stemler, 2001), ilgili çalışma alanına değer vermesi (Butakor, 2016; Eklöf, 2007; Yayan, 2003) ve okula aitlik hissi (Yalçın, Demirtaşlı, Dibek ve Yavuz, 2017) olduğu görülmektedir. Öğrencinin ilgili alanda çalışmayı kolay ya da zor olarak algılamasına bağlı olarak onun o alandaki kendine güven düzeyi ve alana değer vermesi, onun sergileyeceği çabayı ve sonuçta da başarısını etkilemektedir (Akyüz, 2014; Yayan, 2003). Yapılan çalışmalar aynı zamanda öğrencinin belirli bir çalışma alanını sevmesi ile o alandaki başarısı arasında pozitif yönlü ilişkiler olduğunu göstermektedir (Stemler, 2001). Akademik başarıyı etkileyen diğer bir duyuşsal özellik ise okula aitlik hissidir. Okula aitlik hissi, öğrencinin okul ortamları ile sosyal ve duygusal bağlantılarını içerdiğinden akademik başarı açısından önemli bir rol oynamaktadır (Skinner, Furrer, Marchand ve Kindermann, 2008).

Öğrenci çıktıları bağlamında ise en temelde öğrencinin ilgili alandaki akademik başarısı ele alınmaktadır. Bu çalışmada öğrenci öğrenme çıktısı olarak, TIMSS'te ölçüldüğü şekliyle matematik başarısı üzerine odaklanılmıştır. Matematik başarısına odaklanılmasının nedeni ise matematiğin hem bilim, teknoloji ve mühendislik gibi interdisipliner alanlarla ilişkili olması hem de bir ülkede yeni bilimsel ve teknolojik çalışmaların ilerlemesi için önemli bir rol oynamasıdır. Akademik başarının yanısıra, öğrenci çıktıları açısından dikkate alınması gereken faktörlerden biri de algılanan akran baskısıdır. Öğrencilerin akranlarla etkileşimi, onların öğrenme sürecinde önemli rol oynamaktadır. Bu konuda yapılan çalışmalar (Nortvedt, Gustafsson ve W.Lehre, 2016; L. Rutkowski ve D.Rutkowski, 2016), akranlarından daha fazla baskı gören öğrencilerin daha düşük başarı sergilediğini göstermektedir.

Öğretmen kalitesinin, öğretimsel kalitenin ve çeşitli öğrenci niteliklerinin akademik başarı ile ilişkilerinin incelendiği araştırmalarda (Akyüz, 2014; Buluç, 2014; Blömeke ve diğerleri., 2016; Butakor, 2016; Eklöf, 2007; Gustafsson ve Nilsen, 2016; Hong, 2012; House ve Telese, 2008; L. Rutkowski ve D.Rutkowski, 2016; Nortvedt ve diğerleri., 2016; 2011; Stemler, 2001; Yayan, 2003; Yalçın ve diğerleri., 2017; Yoshino, 2012) yaygın olarak yapısal eşitlik modelleme, hiyerarşik lineer modelleme, path (yol) analizi ve regresyon analizlerinin kullanıldığı görülmektedir. Ancak bu konudaki çalışmalarda kümeleme analizi tekniğinden daha az yararlanıldığı gözlenmektedir. Bu analiz teknikleri ile, ilgili özellikler ve başarı arasındaki ilişkiler çalışılabilmektedir. Ancak bu tekniklerden farklı olarak kümeleme analizi tekniğiyle öğrenciler akademik başarı, öğrenci nitelikleri, öğretmen nitelikleri ve öğretimsel nitelikler bakımından sınıflandırılarak öğrenci profilleri

oluşturulabilmektedir. Bu tür bir sınıflamanın ise, matematik başarısı açısından farklı düzeylerde olan öğrencilerin ve onların öğretmenlerinin özellikleri ile bu öğrenciler için matematik dersi öğretim sürecinin nitelikleri konusunda daha detaylı bilgi vereceği düşünülmektedir. Bu nedenle bu araştırmada Nilsen ve diğerleri'nin (2016) önerdiği kavramsal çerçeve temelinde, TIMSS-2015 verilerine dayalı olarak Türk 4. ve 8.sınıf öğrencilerinin; matematik başarısı, öğrenci ve öğretmen nitelikleri ile öğretimsel nitelikler açısından sınıflandırılması amaçlanmıştır. Bu doğrultuda araştırmada öğrenci çıktılarının öğrenci düzeyinde belirleyicileri olarak matematik öğrenmeyi sevme, matematikte kendine güvenme, matematiğe değer verme ve okula aitlik hissi özellikleri ele alınmıştır. Sınıf düzeyinde ise öğretmen kalitesinin göstergeleri olarak öğretmenin öğrenim düzeyi ve mesleki deneyiminin ele alınmasına karar verilmiştir. Öğretimsel kalitenin göstergeleri olarak da öğretmenlerin yanıtlarına dayalı belirlenen, ilgili öğretimsel etkinlikleri gerçekleştirme sıklığı ile öğretmenler arası etkileşimin ve öğrenci yanıtlarına dayalı olarak belirlenen matematik dersinde ilgi çekici öğretime ilişkin görüşlerin ele alınması uygun görülmüştür. Öğrenci çıktıları bağlamında ise matematik başarısı ve algılanan akran baskısı ele alınıp incelenmiştir. Çalışmada bu bağlamda öğrenci ve öğretmen nitelikleri ile öğretimsel nitelikler açısından benzerlik ve farklılıklara dayalı olarak öğrenci profilleri oluşturulmuştur.

Araştırmanın Amacı

Bu araştırma, TIMSS-2015 uygulamasına dayalı olarak, öğrenci çıktıları bağlamında matematik başarısı ve algılanan akran baskısı ile öğretmen kalitesi ve öğretimsel kalitenin göstergeleri ile çeşitli öğrenci nitelikleri açısından dördüncü ve sekizinci sınıf öğrencilerini sınıflandırarak öğrenci profilleri oluşturmak amacıyla gerçekleştirilmiştir.

YÖNTEM

Çalışma Grubu

TIMSS-2015 çalışmasının evrenini Türkiye'de öğrenim görmekte olan dördüncü ve sekizinci öğrencileri ile bu öğrencilerin öğretmenleri oluşturmaktadır. TIMSS-2015 çalışması için örneklem, Milli Eğitim Bakanlığı, Eğitimi Araştırma ve Geliştirme Dairesi Başkanlığı (EARGED) Ölçme ve Değerlendirme Şubesinde kurulan birim tarafından iki aşamalı tabakalı örnekleme yöntemi kullanılarak seçilmiştir. Örneklem seçiminde ilk aşamada okullar, ikinci aşamada ise bu okullardan sınıflar seçkisiz olarak seçilmiştir. Okul seçimleri IEA tarafından gerçekleştirilmiştir. TIMSS-2015 çalışmasına Türkiye'den dördüncü sınıf düzeyinde n=251 öğretmen ve n=6456 öğrenci katılırken, sekizinci sınıf düzeyinde n=220 öğretmen ve n=6079 öğrenci katılmıştır (Yıldırım ve diğerleri., 2016). Ancak bu araştırmaya başlarken, araştırmada incelenen değişkenler açısından kayıp veriler için inceleme yapılmıştır. Buna göre "öğretmenin öğrenim düzeyi" ve "öğretmenin deneyim süresi" değişkenleri açısından kayıp verisi bulunan öğretmenler (4.sınıf düzeyi için n=7 ve 8.sınıf düzeyi için n=4) ve bu öğretmenlerin öğrencileri (dördüncü sınıf düzeyi için n=189 ve sekizinci sınıf düzeyi için n=114) çalışmaya dahil edilmemiştir. Bu öğretmen ve öğrencilerin, TIMSS 2015 çalışmasının örnekleminde yer alan diğer öğretmen ve öğrencilerden, TIMSS 2015 çalışmasında ölçülen özellikler açısından nasıl farklılaştığı bilinmemektedir. Kayıp verisi bulunan bu öğretmen ve öğrencilerin çıkarılması ile bu çalışmada yararlanılan öğretmen ve öğrenci grubu artık TIMSS 2015 örnekleminden farklılaşmış ve örneklem olma niteliğini kaybetmiştir. Bu nedenle bu araştırma için evren ve örneklem tanımı yapılması yerine çalışma grubu tanımlanmıştır. Bu doğrultuda çalışma grubu toplam n=12232 öğrenci ve n=460 öğretmenden oluşmuştur. Çalışma grubundaki öğretmen ve öğrencilerin cinsiyet ve okul açısından dağılımı Tablo 1.'de sunulmuştur.

Tablo 1. Çalışma Grubundaki Öğrencilere İlişkin Bilgiler

Sınıf			N	%	Toplam
4.sınıf	Okul		236	100	236
	Öğretmen	Kadın	140	57.40	244
		Erkek	104	42.60	
	Öğrenci	Kadın	3071	49	6267
		Erkek	3187	50.90	
		Eksik veri	1	0.10	
8.sınıf	Okul		215	100	215
	Öğretmen	Kadın	101	46.76	216
		Erkek	115	53.24	
	Öğrenci	Kadın	2877	48.20	5965
		Erkek	3065	51.40	
		Eksik veri	23	0.40	

Tablo 1.'e bakıldığında çalışma grubunun, 4.sınıf düzeyinden toplam n=244 öğretmen ve n=6267 öğrenci ve sekizinci sınıf düzeyinden n=216 öğretmen ve n=5965 öğrenciden oluştuğu görülmektedir.

Veri Toplama Yöntemleri

Araştırmada kullanılan veri toplama araçları ile ilgili bilgiler bu bölümde belirtilebilir. Başlıktaki üç Araştırmada, TIMSS-2015'de uygulanan matematik başarı testi ile öğrenci ve öğretmen anketlerinde yer alan ölçeklere dayalı olarak elde edilen ölçümler kullanılmıştır. Araştırma amacı doğrultusunda veriler, TIMSS&PIRLS internet sitesinden elde edilmiştir. Öğretmenin deneyim süresi (ATBG01, BTBG01) ve öğrenim düzeyine (ATBG04, BTGB04) ilişkin ölçümler doğrudan TIMSS-2015 uluslararası veri tabanından elde edilmiştir. Ancak "sınıf içi öğretim etkinlikleri" ile "öğretmenler arası etkileşim" değişkenleri için öğretmen anketindeki ve matematikte kendine güvenme, matematik öğrenmeyi sevme, matematik dersinde ilgi çekici öğretime ilişkin görüşler, matematiğe değer verme (sadece 8. sınıf düzeyi için) değişkenleri için ise öğrenci anketlerindeki ilgili maddelere (4'lü Likert türü derecelemeyi kullanan maddeler) verilen yanıtlardan elde edilen puanlara dayalı olarak (4. ve 8. sınıf düzeyi için ayrı ayrı olmak üzere) Açımlayıcı Faktör Analizi (AFA) yoluyla indeks puanları (z puanları) üretilmiştir. Matematik başarısının göstergesi olarak TIMSS-2015 araştırmasında uygulanan matematik başarı testi ile elde edilen ölçümler (Plausible Values; ASMMAT01-05, BSMMAT01-05) kullanılmıştır. AFA öncesinde öğretmenin öğrenim düzeyi ve deneyim süresi değişkenleri dışındaki değişkenler için kayıp veri incelemesi yapılmış, o değişkenler için kayıp veriler aritmetik ortalama ile yer değiştirilmiştir. Öğrenim düzeyi ve deneyim süresi değişkenleri için ise kayıp verisi bulunan öğretmenlere (dördüncü sınıf düzeyi için n=7 ve sekizinci sınıf düzeyi için n=4) ait ve bu öğretmenlerin öğrencilerine (dördüncü sınıf düzeyi için n=189 ve sekizinci sınıf düzeyi için n=114) ait veriler analiz dışı bırakılmıştır.

Bu araştırmada gerçekleştirilen bir seri AFA'da faktör çıkartma tekniği olarak Temel Bileşenler tekniği kullanılmış, faktör sayısına sınırlama konulmamış ve hiçbir döndürme işlemi uygulanmamıştır. Analiz öncesinde ilgili veri setinin analizin varsayımlarını karşılayıp karşılamadığı test edilmiştir. Her bir veri setinin faktör analizine uygunluğunu değerlendirmek üzere Kaiser-Mayer-Olkin (KMO) katsayısı ile Barlett Küresellik Testi sonuçları incelenmiştir. Dördüncü ve sekizinci sınıf düzeylerindeki öğrenciler için ayrı ayrı uygulanan bir seri AFA'ya dayalı olarak elde edilen KMO katsayıları ve Barlett testinde hesaplanan χ^2 değerleri aşağıda Tablo 2.'de sunulmuştur.

Tablo 2. Araştırmada İncelenen Öğrenci-Öğretmen Nitelikleri ve Öğretimsel Niteliklere İlişkin Ölçümlere Yönelik AFA sonuçları ve Cronbach α değerleri

Değişken	Sınıf	Madde	KMO	Barletts χ^2 (sd)	Özdeğer	Açıklanan varyans (%)	Faktör yükleri	α değerleri
Matematik başarısı	4	----------	.93	52278.46** (10)	4.66	93.24	.97	.98
	8	----------	.93	51838.58 ** (10)	4.68	93.55	.97	.98
Matematikte Kendine güvenme	4	MS3A-MS3I	.87	15984.61** (36)	3.72	41.28	-.63- .73	.53
	8	19A-19I	.89	25729.82** (36)	4.50	50.01	-.72- .50	.59
Matematik öğrenmeyi sevme	4	MS1A-MS1I	.91	22086.58** (36)	4.41	48.94	-.66- .80	.61
	8	17A-17I	.93	34797.38** (36; .00)	5.47	60.82	-.73- .89	.64
Matematik dersinde ilgi çekici öğretime ilişkin görüşler	4	MS2A-MS2J	.91	15047.44** (45)	3.87	38.71	.42- .71	.79
	8	18A-18J	.94	26970.79** (45)	5.20	52.01	.52- .80	.89
Matematiğe değer verme	8	20A-201	.90	22688.60** (36)	4.54	50.48	.56- .81	.87
Okula aitlik hissi	4	G11A-G11G	.84	7114.24** (21)	2.73	38.90	.37- .69	.70
	8	15A-15G	.86	92229.39** (21)	3.06	43.78	.51- .74	.78
Akran baskısı	4	G12A-G12H	.89	11815.68** (28)	3.43	42.88	.58- .72	.80
	8	16A-16I	.89	14941.77** (36)	3.89	43.28	.55- .76	.82
Sınıf içi öğretim etkinlikleri	4	15A-15H	.79	9641.71** (28)	2.89	36.2	.43- .76	.73
	8	15A-15G	.74	6308.79** (21)	2.51	35.88	.52- .67	.70
Öğretmenler arası etkileşim	4	10A-10G	.88	18984.28** (21)	3.90	55.66	.56- .81	.86
	8	10A-10G	.86	16405.85** (21)	3.77	53.84	.68- .82	.86

**p<.01

Tablo 2.'de sunulan KMO katsayılarının .80'den yüksek olması (sınıf içi öğretim etkinlikleri ölçümleri için hesaplanan KMO değerleri .80'den düşüktür ancak bu değerler de .80'e yakın oldukları için bu veri setlerinin de faktörleştirilebilir nitelikte olduğu sonucuna varılmıştır), her bir veri setinin faktörleştirilebilir nitelikte olduğuna işaret etmektedir. Hesaplanan Barlett χ^2 değerlerinin manidar olması ise, her bir veri setinin açımlayıcı faktör analizi için uygun olduğunu göstermektedir. Bu özelliklere ilişkin ölçümler için tek faktörlü bir yapının elde edilmesi, açıklanan varyans oranlarının yüksek olması ve faktör yüklerinin .32'den büyük olması, söz konusu ölçümlerin ilgili yapıların uygun birer temsilcisi olduklarına işaret etmektedir ve bu bulgular söz konusu ölçümlerin yapı geçerliğine ilişkin kanıtlar olarak değerlendirilmiştir (Büyüköztürk, 2012). İlgili ölçümlerin

güvenirliğine ilişkin kanıtlar elde etmek üzere hesaplanan Cronbach α değerleri incelendiğinde ise bazı ölçümler için düşük olmakla birlikte genel olarak söz konusu maddelerin ilgili yapının kabul edilebilir düzeyde güvenilir ölçümlerini sağladığına işaret etmektedir.

Verilerin Analizi

Araştırmanın başında kayıp veriler ile bunların analizler için bir sorun teşkil edip etmeyeceğini incelemek amacı ile Little's MCAR (Missing Completely at Random) testi yapılmıştır (Öğretmenin öğrenim düzeyi ve deneyim süresi dışındaki değişkenler için uygulanmıştır). Dördüncü ve sekizinci sınıf öğrenci anketlerinden elde edilen veri setleri için ayrı ayrı yapılan testler sonucunda elde edilen bulgular (4.sınıf için χ^2=16776.51, p<.000; 8.sınıf için χ^2=18722.39, p<.000), bu veri setlerinin rastgele örüntüler içerdiğini, öğretmen anketlerinden elde edilen veri setleri için yapılan testler sonucunda ulaşılan bulgular ise (4.sınıf için χ^2=55.93, p>.000; χ^2=26.12, p>.000), kayıp verilerin herhangi bir örüntü içermediğini göstermektedir (Garson, 2015). Bu doğrultuda her iki sınıf düzeyinde ilgili veri setleri için kayıp verinin analizler için bir sorun teşkil etmeyeceği sonucuna ulaşılmış ve kayıp veriler ilgili değişken için aritmetik ortalama değerleriyle yer değiştirilmiştir. Öğretmenin öğrenim düzeyi ve deneyim süresi değişkenleri için kayıp verileri bulunan bireyler çalışma grubuna dahil edilmemiştir. Bu araştırma öğrenci çıktıları bağlamında matematik başarısı ve algılanan akran baskısı ile öğretmen kalitesi ve öğretimsel kalitenin göstergeleri ile çeşitli öğrenci nitelikleri açısından öğrencileri sınıflandırmak amacıyla gerçekleştirilmiştir. Bu amaç doğrultusunda araştırmada kümeleme analizi tekniği uygulanmıştır. Kümeleme analizi ilgilenilen özellikler açısından veri setindeki her bir gözlemi, her bir grup (küme) içerisindeki gözlemlerin birbirine benzer, ancak grupların birbirlerinden farklı olmasını sağlayacak şekilde gruplara atamak amacıyla gerçekleştirilmektedir. Kümeleme analizinde çeşitli yöntemler kullanılmaktadır (Özdamar, 2014). Bu araştırmada ise hiyerarşik ve hiyerarşik olmayan kümeleme algoritmalarını birleştiren bir yöntem olan iki-aşamalı kümeleme yöntemi kullanılmıştır. Bu yöntemin, tek bir kümeleme algoritması kullanılmasının sınırlılıklarını azalttığı ve sayısal ve kategorik verileri içeren veri setleri için bu yöntemle daha sağlam/güçlü sonuçların elde edilebildiği belirtilmektedir (Shih, Jheng ve Lai, 2010).

Analiz öncesinde kümeleme analizi tekniğinin varsayımlarının 4.sınıf ve 8.sınıf düzeyinde, ilgili veri setleri için karşılanıp karşılanmadığı incelenmiştir. Kümeleme analizinin en temel varsayımı, ölçümlerin geçerli olmasıdır. Bu araştırmada kullanılan ölçümlerin geçerliğine ilişkin kanıtlar "Veriler" bölümünde sunulmaktadır ve bu kanıtlara dayalı olarak bu varsayımın karşılandığı düşünülmektedir. Kümeleme analizi için diğer varsayım ise, değişkenler arası çoklu bağlantı sorununun olmamasıdır. Bu doğrultuda 4. ve 8.sınıf düzeyi için araştırmadaki değişkenler arasındaki ilişkilere yönelik korelasyonlar ayrı ayrı hesaplanmıştır. Öğretmenin öğrenim düzeyi ile diğer değişkenler arasındaki ilişkiler Spearman sıra farkları korelasyon tekniği ile diğer tüm değişkenler arasındaki ilişkiler ise Pearson momentler çarpımı korelasyon tekniği ile incelenmiştir. Hesaplanan korelasyon katsayıları aşağıda Tablo 3.'te sunulmuştur.

Tablo 3. Dördüncü ve Sekizinci Sınıf Düzeyleri için Değişkenler Arası İlişkiler

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
x1	1.00	-.50*	.27*	.32*	.27*	-.31*	.13*	.05*	.31*	-.10*	
x2	-.49*	1.00	-.52*	-.34*	-.27*	.29*	-.03*	-.03*	-.11*	.04*	
x3	.19*	-.70*	1.00	.52*	.47*	-.28*	.00	-.00	.05*	-.03*	
x4	.12*	-.33*	.51*	1.00	.54*	-.32*	.04*	-.01	.08*	.00	
x5	-.02	-.15*	.36*	.45*	1.00	-.34*	.04*	-.01	.09*	-.03*	
x6	-.19*	.09*	-.06*	-.15*	-.27*	1.00	-.04*	.01	-.12*	.02	
x7	.09*	-.01	-.03*	.01	-.03*	-.02	1.00	.32*	.10*	.06*	
x8	.06*	.01	-.05*	-.03*	-.06*	-.01	.42*	1.00	-.05*	.05*	
x9	.23*	-.02	-.08*	.10*	-.09*	-.03*	.05*	.08*	1.00	-.48*	
x10	.12*	-.02	.00	.03*	-.03*	.01	.07*	.05*	-.11*	1.00	

x11	.15*	-.42*	.60*	.51*	.35*	-.07*	.00	-.03*	-.04*	-.02	1.00

*p<.05

Not: x1:Matematik başarısı, x2:Matematikte kendine güvenme, x3:Matematik öğrenmeyi sevme, x4:Matematik dersinde ilgi çekici öğretime ilişkin görüşler, x5:Okula aitlik hissi, x6:Akran baskısı, x7:Sınıf içi öğretim etkinlikleri, x8: Öğretmenler arası etkileşim, x9: Öğretmenin deneyim süresi, x10: Öğretmenin öğrenim düzeyi, x11: Matematiğe değer verme (Matriste üst diagonal 4. Sınıf düzeyi için, alt diagonal ise 8. Sınıf düzeyi için hesaplanan korelasyon katsayılarını göstermektedir).

Tablo 3.'te sunulan korelasyon katsayılarına bakıldığında r=.80'in üzerinde bir korelasyon katsayısının bulunmadığı görülmektedir ve dolayısıyla "çoklu bağlantı" sorununun olmadığı anlaşılmaktadır. Varsayımların test edilmesinin ardından, kümeleme analizi yapılmıştır.

BULGULAR

Araştırma amacı doğrultusunda dördüncü ve sekizinci sınıf düzeyi için ilgili very setleri üzerinde -küme sayısına ilişkin bir sınırlama yapılmaksızın- iki aşamalı kümeleme analizi gerçekleştirilmiştir. Dördüncü sınıf düzeyi için gerçekleştirilen iki-aşamalı kümeleme analizi sonucunda öğrencilerin üç kümeye ayrıştığı gözlenmiştir. Ardından bu kümeleme yapısının geçerliğini incelemek üzere farklı küme sayılarına ilişkin ortalama Silhouette genişliği (Silhouette width) hesaplanmıştır. İki küme için ortalama silhouette genişliği .54, üç küme için .60, dört küme için .48 ve beş küme için .46 olarak hesaplanmıştır. Hesaplanan bu değerler, dördüncü sınıf düzeyindeki öğrenciler için yapılan kümeleme işlemindeki en uygun küme sayısının üç olduğuna işaret etmektedir (Rousseuw, 1987). Bu doğrultuda elde edilen üç kümeye ilişkin sonuçlar Tablo 4.'te sunulmuştur.

Tablo 4. Dördüncü Sınıf Öğrencilerinden Elde Edilen Veri Seti İçin İki Aşamalı Kümeleme Analizi Sonuçları

Küme	N	%
1	3032	48.40
2	2049	32.70
3	1186	18.90

Tablo 4'e bakıldığında, çalışma grubunda yer alan dördüncü sınıf düzeyindeki öğrencilerin üç farklı kümeye ayrıştığı gözlenmiştir. Çalışma grubundaki 4.sınıf düzeyindeki öğrencilerin yaklaşık yarısı (%48.40) 1.kümede yer almıştır. Üçüncü kümenin ise en küçük küme olduğu ve çalışma grubundaki 4.sınıf düzeyindeki öğrencilerin sadece yaklaşık %19'unun bu kümede yer aldığı görülmektedir.

Kümelerin belirlenmesinin ardından, araştırmada ele alınan her bir değişken için yordayıcı önem değerleri hesaplanmıştır. Buna göre öğrenci özelliklerinden "matematik başarısı", "matematikte kendine güvenme", "matematik öğrenmeyi sevme", "algılanan akran baskısı" ile öğretmene ilişkin özelliklerden "deneyim süresi" ve "öğrenim düzeyi"nin bu kümelerin oluşmasına en fazla katkı sağlayan (yordayıcı önem değeri=1.00) değişkenlerdir. Öğrencilerin matematik dersinde ilgi çekici öğretime ilişkin görüşleri (yordayıcı önem değeri=.82) ile okula aitlik hislerinin de (yordayıcı önem değeri=.79), öğrencilerin sınıflandırılmasında önemli katkıları olduğu görülmüştür. Ancak sınıf için öğretim etkinlikleri değişkeninin (yordayıcı önem değeri=.07) öğrencilerin sınıflanmasında çok az önemli olduğu, öğretmenler arası etkileşim düzeyinin (yordayıcı önem değeri=.01) ise neredeyse hiçbir öneminin olmadığı gözlenmiştir. Bu durum, öğretmenlerin ilgili öğretimsel etkinlikleri gerçekleştirme sıklığı ve birbirleriyle olan etkileşim düzeyleri açısından çalışma grubundaki 4.sınıf öğrencilerinin benzer yapıda olduklarına işaret etmektedir.

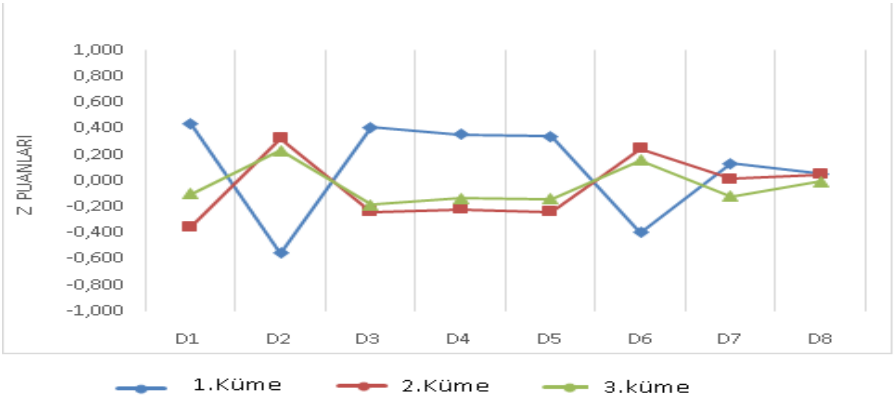
Ardından, bu üç kümenin, araştırmada ele alınan değişkenler açısından özellikleri ayrıntılı olarak incelenmiştir. Bu inceleme sonucunda; birinci kümedeki öğrencilerin tamamı (n=3032, %100) ile

ikinci kümedeki öğrencilerin neredeyse tamamının (n=2041, %99.60) öğretmenlerinin lisans düzeyinde öğrenim görmüş oldukları, üçüncü kümedeki öğrencilerin ise büyük bir kısmının (n=991, %83.60) öğretmenlerinin ön-lisans (Yüksekokul-2 yıllık) düzeyinde öğrenim görmüş olduğu ve küçük bir kısmının (n=195, %16.40) ise öğretmenlerinin lisansüstü düzeyde öğrenim görmüş olduğu anlaşılmaktadır. Araştırmada ele alınan diğer değişkenlerin hepsi sürekli değişkenlerdir. Bu nedenle, ilgili değişkenler açısından bu üç kümedeki öğrencilerin durumunu ortaya koymak üzere, ilgili veri setlerine ilişkin betimsel istatistikler hesaplanarak Tablo 5.'te sunulmuştur.

Tablo 5. Dördüncü Sınıf Düzeyinde İki-Aşamalı Kümeleme Analizi İle Elde Edilen Kümeler İçin Araştırmada İncelenen Değişkenlere İlişkin Betimsel İstatistikler

Değişken	**1.küme**		**2.küme**		**3.küme**	
	$\overline{X}$	$(SS)^2$	$\overline{X}$	$(SS)^2$	$\overline{X}$	$(SS)^2$
Matematik başarısı	.44	.51	-.79	.87	.25	.84
Matematikte kendine güvenme	-.55	.41	.88	.68	-.09	.98
Matematik öğrenmeyi sevme	.41	.14	-.65	1.73	.06	.82
Matematik dersinde ilgi çekici öğretime ilişkin görüşler	.36	.14	-.58	1.96	.09	.66
Okula aitlik hissi	.34	.20	-.58	1.91	.10	.65
Algılanan akran baskısı	-.40	.39	.64	1.38	-.09	.81
Sınıf içi öğretim etkinlikleri	.13	.79	-.12	1.06	-.13	1.34
Öğretmenler arası etkileşim	.06	1.05	-.01	.97	-.05	.82
Öğretmenin deneyim süresi	15.01	67.81	11.31	63.55	28.35	97.69

Öğretmenin deneyim süresine ilişkin ölçümler dışındaki tüm değişkenlere ilişkin ölçümler AFA yoluyla elde edilen z puanlarıdır. Bu doğrultuda öğretmenin deneyim süresi dışındaki değişkenler için betimsel istatistiklere dayalı olarak, ilgili değişken açısından o kümedeki öğrencilerin ve öğretmenlerin, çalışma grubundaki 4.sınıf düzeyindeki öğrencilerin ve öğretmenlerin tamamı ile karşılaştırmak mümkün olmaktadır. Buna göre matematik başarısı, matematik öğrenmeyi sevme, matematik dersinde ilgi çekici öğretime ilişkin görüşler ve okula aitlik hissi değişkenleri açısından birinci ve üçüncü kümedeki öğrenciler, çalışma grubunda 4.sınıf düzeyindeki öğrencilerin ortalamasından yüksek düzeydedirler. İkinci kümedeki öğrencilerin ise matematik dersinde kendine güvenme ve algılanan akran baskısı değişkenleri açısından, çalışma grubunda 4.sınıf düzeyindeki öğrencilerin ortalamasından yüksek ancak diğer değişkenler açısından ortalamadan düşük düzeyde oldukları gözlenmiştir. Her bir küme içerisinde bireysel farklılıklar bağlamında bakıldığında; matematikte kendine güvenme, sınıf içi öğretim etkinlikleri ve deneyim süresi dışındaki değişkenler açısından en fazla değişkenliğin ikinci kümede olduğu anlaşılmaktadır. Ayrıca her bir kümedeki öğrencilerin, araştırmada ele alınan değişkenler açısından durumunu görsel olarak incelemek ve hem bu incelemeye hem de betimsel istatistiklere dayalı olarak öğrenci profilleri oluşturmak üzere aşağıda şekil 1'de sunulan grafik oluşturulmuştur (Öğretmenin deneyim süresi farklı bir ölçekle ölçüldüğünden grafikte bu değişkene yer verilmemiştir).



Şekil 1. Dördüncü Sınıf Düzeyinde Elde Edilen Kümelerdeki Öğrencilerin Araştırmada İncelenen Değişkenlere İlişkin Durumu

Not: D1: Matematik başarısı, D2: Matematikte kendine güvenme, D3:Matematik öğrenmeyi sevme, D4: Matematik dersinde ilgi çekici öğretime ilişkin görüşler, D5: Okula aitlik hissi, D6: Akran baskısı, D7: Sınıf içi öğretim etkinlikleri, D8: Öğretmenler arası etkileşim

Bu üç kümenin her birinde yer alan öğrencilere ve onların öğretmenlerine ilişkin özellikler detaylı bir şekilde incelenerek üç ayrı öğrenci profili oluşturulmuştur:

1. *Küme- Matematik başarısı yüksek ancak matematikte kendine çok az güvenen öğrenciler:* Bu küme, çalışma grubunda 4.sınıf düzeyindeki öğrenciler içerisinde matematik başarısı açısından en yüksek ancak matematikte kendine güvenme açısından en düşük düzeyde olan öğrencilerden oluşmaktadır. Bu kümedeki öğrencilerin üç küme içerisinde; matematik öğrenmeyi en fazla seven, matematik dersinde ilgi çekici öğretim yapıldığını düşünen ve okula aitlik hissine en fazla sahip olan öğrenciler oldukları anlaşılmaktadır. Ayrıca bu kümede algılanan akran baskısı açısından, çalışma grubunda 4. sınıf düzeyindeki öğrencilerin ortalamasından düşük düzeyde ($\overline{X}$=-.40) öğrencilerin bulunduğu gözlenmiştir. Bun yanısıra bu kümedeki öğrencilerin öğretmenlerinin tamamının lisans düzeyinde öğrenim görmüş, çalışma grubunda yer alan 4.sınıf öğretmenleri içerisinde TIMSS-2015 çalışmasında belirtilen öğretimsel etkinlikleri en sık uygulayan, öğretmenler arası etkileşimi en fazla olan ve ortalama 15 yıllık deneyime sahip öğretmenler oldukları anlaşılmaktadır.
2. *Küme- Matematikte kendine çok güvenen ancak matematik başarısı açısından düşük düzeyde olan öğrenciler*: Bu küme, çalışma grubunda 4.sınıf düzeyindeki öğrenciler içerisinde matematikte kendine en fazla güvenen ancak matematik başarısı açısından en düşük düzeyde olan öğrencilerden oluşmaktadır. Üç kümedeki öğrenciler karşılaştırıldığında bu kümedeki öğrencilerin; matematiği en az seven, matematik dersinde ilgi çekici öğretime ilişkin görüşler ($\overline{X}$=-.58) ve okula aitlik hissi açısından en düşük ortalamaya ($\overline{X}$=-.58) sahip, akran baskısına en fazla maruz kalan öğrenciler oldukları görülmektedir. Sınıf düzeyindeki nitelikler açısından incelendiğinde; bu kümedeki öğrencilerin öğretmenlerinin tamamına yakınının (%99.60) lisans düzeyinde öğrenim görmüş, ilgili öğretimsel etkinlikleri daha az uygulayan ($\overline{X}$=-.12), öğretmenler arası etkileşim açısından da ortalamaya çok yakın düzeyde olan ($\overline{X}$=-.01) ve en düşük deneyime sahip (ortalama 11 yıl) öğretmenler oldukları anlaşılmaktadır.
3. *Küme: Matematik başarısı açısından ortalamadan yüksek düzeyde olan ve matematik öğrenmeyi seven öğrenciler:* Bu küme, çalışma grubunda 4. sınıf düzeyindeki öğrencilerin ortalaması ile karşılaştırıldığında; matematik başarısı açısından ortalamadan yüksek ($\overline{X}$=.25), matematiği sevme ($\overline{X}$= .06), matematik dersinde ilgili çekici öğretime ilişkin görüşler ($\overline{X}$=.09) ve okula aitlik hissi ($\overline{X}$=.10) açısından ortalamaya yakın düzeyde olan ancak algılanan akran baskısı ($\overline{X}$= -.09) açısından ortalamadan biraz düşük düzeyde olan öğrencilerden oluşmaktadır. Bununla birlikte bu kümedeki öğrencilerin öğretmenlerinin en fazla deneyime sahip ancak büyük bir kısmının (%83.60) ön lisans düzeyinde öğrenim görmüş, çalışma grubunda yer alan 4.sınıf öğretmenleri içerisinde TIMSS-2015 çalışmasında belirtilen öğretimsel etkinlikleri en az sıklıkta uygulayan, öğretmenler arası etkileşimi en düşük düzeyde olan öğretmenler oldukları anlaşılmaktadır.

Dördüncü sınıf düzeyindeki öğrencilere ilişkin kümeleme analizi sonrasında, sekizinci sınıf düzeyindeki öğrenciler ve onların matematik öğretmenlerine ilişkin veri setleri üzerinde iki-aşamalı kümeleme analizi gerçekleştirilmiştir. Yapılan analiz sonucunda, çalışma grubunda yer alan sekizinci sınıf düzeyindeki öğrencilerin iki kümeye ayrıştığı gözlenmiştir. Sonrasında farklı küme sayıları için hesaplanan ortalama Silhouette genişliği (iki küme için 0.59, üç küme için 0.46, dört küme için 0.36 ve beş küme için 0.34), bu öğrenciler için yapılan kümeleme işlemindeki en uygun küme sayısının 2 olduğuna işaret etmektedir (Rousseeuw, 1987). Bu elde edilen 2 kümeye ilişkin sonuçlar Tablo 6.'da sunulmuştur.

Tablo 6. Sekizinci Sınıf Öğrencilerinden Elde Edilen Veri Seti İçin İki Aşamalı Kümeleme Analizi Sonuçları

Küme	N	%
1	3041	51
2	2924	49

Tablo 6'da görüldüğü üzere, çalışma grubundaki 8.sınıf düzeyindeki öğrencilerin yarısı (%51) 1.kümede ve diğer yarısı ise 2.kümede yer almıştır. Bu kümelerin belirlenmesinin ardından araştırmadaki her bir değişken için yordayıcı önem değerleri hesaplanmıştır. Buna göre öğrenci özelliklerinden "matematiğe verilen değer", "matematikte kendine güvenme"nin, "matematik dersinde ilgi çekici öğretime ilişkin görüşler"in ve "matematiği sevme"nin bu kümelerin oluşturulmasına en fazla katkı sağlayan (yordayıcı önem değeri=1.00) değişkenlerdir. Öğrencilerin okula aitlik hislerinin de (yordayıcı önem değeri=.69), bu sınıflamada önemli katkıları olduğu gözlenmiştir. Öğretmenin öğrenim düzeyi (yordayıcı önem değeri=.35), algılanan akran baskısı (yordayıcı önem değeri=.34) ile matematik başarısının (yordayıcı önem değeri=.28) ise bu kümeleme işleminde daha az katkılarının olduğu bulunmuştur. Öğretmenin deneyim süresi (yordayıcı önem değeri=.06), sınıf içi öğretimsel etkinlikler (yordayıcı önem değeri=.03) ve diğer öğretmenlerle etkileşim değişkenlerine (yordayıcı önem değeri=.01) ilişkin hesaplanan yordayıcı önem değerleri ise, öğrencilerin bu şekilde sınıflanmasında bu özelliklerin çok az önemli olduklarına ya da hiç önemli olmadıklarına işaret etmektedir. Bu bulgular, bu özellikler açısından bu iki kümedeki öğrencilerin benzer yapıda oldukları şeklinde yorumlanabilir.

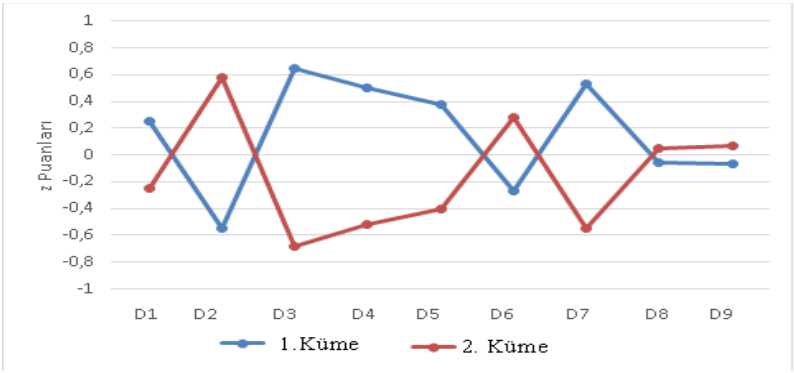
Ardından, bu iki kümede yer alan öğrenciler ile onların öğretmenlerinin özellikleri ayrıntılı olarak incelenmiştir. Birinci kümedeki öğrencilerin neredeyse tamamının (n=3035, %99.80) öğretmenleri lisans düzeyinde, ikinci kümedeki öğrencilerin ise büyük bir kısmının (n=2459, %84.10) öğretmenleri lisans düzeyinde ve çok az bir kısmının (n=153, %5.20) ön-lisans (Yüksekokul-2 yıllık) düzeyinde ve yine çok az bir kısmının (n=22, %0.80) öğretmenleri ise lisansüstü düzeyde öğrenim görmüştür. Ayrıca araştırmada ele alınan diğer değişkenler açısından bu iki kümedeki öğrenciler ve onların öğretmenlerine ilişkin veri setleri için betimsel istatistikler hesaplanarak Tablo 7'de sunulmuştur.

Tablo 7. Sekizinci Sınıf Düzeyinde İki-Aşamalı Kümeleme Analizi İle Elde Edilen Kümeler İçin Araştırmada İncelenen Değişkenlere İlişkin Betimsel İstatistikler

Değişken	1.küme		2.küme	
	$\overline{X}$	$(SS)^2$	$\overline{X}$	$(SS)^2$
Matematik başarısı	.25	1.02	-.25	.85
Matematikte kendine güvenme	-.55	.70	.58	.67
Matematik öğrenmeyi sevme	.65	.29	-.68	.84
Matematik dersinde ilgi çekici öğretime ilişkin görüşler	.50	.20	-.52	1.30

Matematiğe değer verme	.53	.27	-.55	1.16
Okula aitlik hissi	.38	.45	-.40	1.28
Algılanan akran baskısı	-.27	.40	.28	1.46
Sınıf içi öğretim etkinlikleri	-.06	1.06	.05	.96
Öğretmenler arası etkileşim	-.07	.96	.07	1.05
Öğretmenin deneyim süresi	8.73	5.41	1.62	84.96

Bu istatistikler incelendiğinde 1.kümedeki öğrencilerin; matematik başarısı, matematik öğrenmeyi sevme, matematik dersinde ilgi çekici öğretime ilişkin görüşler, matematiğe değer verme ve okula aitlik hissi açısından, çalışma grubunda yer alan 8.sınıf düzeyindeki öğrencilerin ortalamasından yüksek düzeyde oldukları anlaşılmaktadır. İkinci kümedeki öğrencilerin ise, matematikte kendine güvenme ve algılanan akran baskısı değişkenleri açısından çalışma grubunda yer alan 8.sınıf düzeyindeki öğrencilerin ortalamasından yüksek düzeyde olmakla birlikte diğer değişkenler açısından ortalamadan düşük düzeyde oldukları görülmektedir. Hesaplanan varyans değerleri incelendiğinde 1.kümede matematik başarısı, sınıf içi öğretim etkinlikleri, matematikte kendine güvenme değişkenleri açısından bireysel farklılıkların daha fazla olduğu; 2.kümede ise matematik öğrenmeyi sevme, matematik dersinde ilgi çekici öğretime ilişkin görüşler, matematiğe değer verme, algılanan akran baskısı, öğretmenler arası etkileşim ve öğretmenin deneyim süresi değişkenleri açısından bireysel farklılıkların daha fazla olduğu anlaşılmaktadır. Ardından, bu iki kümedeki öğrencilerin araştırmada ele alınan değişkenler açısından durumunu görsel olarak ortaya koymak üzere bir grafik oluşturulmuş ve bu grafik aşağıda şekil 2.'de sunulmuştur.



Şekil 2. Sekizinci Sınıf Düzeyinde Elde Edilen Kümelerdeki Öğrencilerin Araştırmada İncelenen Değişkenlere İlişkin Durumu

Not: D1: Matematik başarısı, D2: Matematikte kendine güvenme, D3: Matematik öğrenmeyi sevme, D4: Matematik dersinde ilgi çekici öğretime ilişkin görüşler, D5: Okula aitlik hissi, D6: Akran baskısı, D7: Matematiğe değer verme, D8: Sınıf içi öğretim etkinlikleri, D9: Öğretmenler arası etkileşim

Bu iki kümede yer alan öğrenciler ve onların matematik öğretmenlerine ilişkin özellikler ayrıntılı bir şekilde incelenmiş ve buna göre iki ayrı öğrenci profili oluşturulmuştur:

1. *Küme- Matematik başarısı açısından iyi düzeyde olan ancak matematikte kendine çok az güvenen öğrenciler:* Bu kümenin matematik başarısı açısından, çalışma grubunda 8. sınıf düzeyindeki öğrencilerin ortalamasından yüksek düzeyde ($\overline{X}$=.25) olan ancak matematikte kendilerine çok az güvenen öğrencilerden oluştuğu görülmektedir. Bu kümede aynı zamanda matematiği çok seven, matematik dersinin ilgi çekici nitelikte olduğunu düşünen, matematiğe çok değer veren, okula aitlik hissi açısından çalışma grubunda 8.sınıf düzeyinde yer alan öğrencilerin ortalamasından yüksek düzeyde ($\overline{X}$=.38) olan öğrenciler bulunmaktadır. Algılanan akran baskısı açısından ise bu kümedeki öğrencilerin, çalışma grubunda 8.sınıf düzeyinde yer alan öğrencilerin

ortalamasından düşük düzeyde ($\overline{X}$=-.27) oldukları gözlenmiştir. Sınıf düzeyindeki nitelikler açısından bakıldığında; bu kümedeki öğrencilerin öğretmenlerinin neredeyse tamamının (%99.80) lisans düzeyinde öğrenim görmüş, öğretimsel etkinlikleri uygulama sıklığı ($\overline{X}$=-.06) ve öğretmenler arası etkileşim düzeyi ($\overline{X}$=-.07) açısından ortalamadan biraz düşük düzeyde olan ve ortalama yaklaşık 9 yıllık deneyime sahip öğretmenler oldukları anlaşılmaktadır.

2. Küme-Matematikte kendine çok güvenen ancak matematik başarısı açısından düşük düzeyde olan öğrenciler: Bu küme, matematikte kendine çok güvenen ancak matematik başarısı açısından çalışma grubunda 8.sınıf düzeyindeki öğrencilerin ortalamasından düşük düzeyde ($\overline{X}$=-.25) olan öğrencilerden oluşmaktadır. Bu kümede aynı zamanda matematiği sevme ($\overline{X}$=-.68), matematik dersinde ilgi çekici öğretime ilişkin görüşler ($\overline{X}$=-.52), matematiğe değer verme ($\overline{X}$=-.55) ve okula aitlik hissi ($\overline{X}$=-.40) açısından çalışma grubunda 8.sınıf düzeyindeki öğrencilerin ortalamasından düşük düzeyde olan öğrenciler bulunmaktadır. Bu öğrencilerin, 1.kümedeki öğrencilere göre, daha fazla akran baskısına maruz kaldıkları gözlenmiştir. Bunun yanısıra bu kümede yer alan öğrencilerin büyük bir kısmının (%84) öğretmenlerinin lisans düzeyinde öğrenim görmüş, öğretimsel etkinlikleri uygulama sıklığı ($\overline{X}$=.05) ve öğretmenler arası etkileşim düzeyi ($\overline{X}$=.07) açısından ortalamaya yakın düzeyde olan ve ortalama yaklaşık 11 yıllık deneyime sahip öğretmenler oldukları anlaşılmaktadır.

SONUÇLAR ve TARTIŞMA

Bu araştırmada TIMSS-2015 uygulamasına dayalı olarak, Türk 4. ve 8.sınıf öğrencilerinin matematik başarısı, öğrenci ve öğretmen nitelikleri ile öğretimsel nitelikler açısından sınıflandırılarak öğrenci profilleri oluşturulması amaçlanmıştır. Bu bağlamda burada esas alınan kuramsal çerçeve temelinde öğretmen kalitesi, öğretimsel kalite ve öğrenci niteliklerinin öğrenci başarısını nasıl ve ne yönde ilişkili olduğu da incelenmiştir. Bu tür bir incelemenin, matematik başarısı açısından öğrencilerin farklılaşmasına neden olabilecek öğrenci ve öğretmen nitelikleri ile öğretimsel niteliklerin ortaya çıkarılmasını sağladığı düşünülmektedir.

Araştırmada uygulanan analizler sonucunda dördüncü sınıf düzeyinde öğrencilerin üç kümeye, sekizinci sınıf düzeyinde ise iki kümeye ayrıştığı gözlenmiştir. Analiz sonuçları, dördüncü sınıf öğrencileri için yapılan sınıflamada öğrenci düzeyinde en etkili özelliklerin, öğrencilerin matematik başarısı, matematikte kendine güvenme, matematik öğrenmeyi sevme ve okula aitlik hissi olduğunu göstermektedir. Sekizinci sınıf öğrencileri için ise, öğrenci düzeyinde ele alınan özelliklerden matematiğe değer verme, matematikte kendine güvenme, matematik öğrenmeyi sevme ve okula aitlik hissinin öğrencilerin sınıflanmasında en önemli özellikler olduğu bulunmuştur. Ancak bu sınıf düzeyindeki öğrencilerin kümelere ayrıştırılmasında matematik başarısının düşük düzeyde etkisinin olduğu gözlenmiştir. Bu bulgular öncelikle matematik başarısı açısından, dördüncü sınıf düzeyinde ortaya çıkan bu kümelerdeki öğrencilerin önemli ölçüde farklılaştığına ancak sekizinci sınıf düzeyinde elde edilen kümelerdeki öğrenciler için bu kadar belirgin bir farklılığın ortaya çıkmadığına işaret etmektedir. Bu durumun, öğrencilerin gelişimsel özellikleri ile ilişkili olduğu düşünülmektedir. Psikososyal gelişim açısından değerlendirildiğinde alanyazında bu bulguyu destekleyecek şekilde, bu dönemdeki çocukların sosyal ortamlarda ve özellikle okul ortamında başarıları aracılığıyla kendilerini göstermek istedikleri ve başarılarına dayalı olarak takdir edilmenin onlar için önemli olduğu belirtilmektedir. Sekizinci sınıf öğrencilerinin sınıflanmasında ise matematik başarısından ziyade duyuşsal özelliklerin ön plana çıktığı görülmektedir. Bu bulgu, ergenlik döneminde olan bu öğrencilerin psiko-sosyal gelişim özellikleriyle örtüşmektedir. Ergenliğe geçişle birlikte bireysel farklılıklar daha fazla önemsenmekte, kişi kendini özgün ayırıcı yönleri ile ortaya koymakta ve kendini daha çok içsel özellikleri ile tanımlamaya başlamaktadır. Alanyazında ergenlik döneminde duyguların yoğunluğundaki artış ile birlikte bu dönemdeki kişilerin davranış biçimlerinde ve çevre ile olan ilişkilerinde duyguların yönlendirici nitelikte olduğu belirtilmektedir (Erikson, 1968). Analiz sonuçları ayrıca, her iki sınıf düzeyinde elde edilen kümelerdeki öğrencilerin matematik öğrenmeyi sevme, matematikte kendine güvenme ve okula aitlik hissi açısından önemli ölçüde farklı olduklarını ortaya koymaktadır.

Sınıf düzeyinde ise öğretmen niteliklerinden öğrenim düzeyinin ve deneyim süresinin, dördüncü sınıf öğrencilerinin kümelere ayrıştırılmasında çok fazla katkısı olduğu görülmüştür. Bu sonuçlar, ilgili alanyazındaki sonuçlarla tutarlılık göstermektedir. Blömeke ve diğerleri (2016) tarafından TIMSS-2011 uygulamasına dayalı olarak dördüncü sınıf düzeyi için uluslararası boyutta gerçekleştirilen çalışmada öğretmenin öğrenim düzeyinin, tüm ülkeler arası öğrenci başarısının en güçlü yordayıcısı olduğu bulunmuştur. İlgili araştırmada ayrıca öğretmenin deneyim süresinin de öğrenci çıktılarındaki bireysel farklılıkları açıklamada manidar düzeyde katkısının olduğu gözlenmiştir. Öğretmenin deneyim süresinin öğrenciyi tanıma, matematik alan bilgisi ve genel olarak eğitim alanındaki bilgisinin artmasına yol açmasıyla birlikte öğrenci çıktıları üzerinde etkili olduğu düşünülmektedir. Ancak sekizinci sınıf düzeyinde, öğrencilerin sınıflanmasında öğretmen niteliklerinden öğrenim düzeyinin düşük düzeyde etkili olmakla bitlikte deneyim süresinin neredeyse hiçbir etkisinin bulunmadığı gözlenmiştir. Bu bulgu, sekizinci sınıf düzeyinde elde edilen iki kümedeki öğrencilerin öğretmenlerinin eğitimsel geçmiş açısından benzer niteliklere sahip olduğunu göstermektedir. Bu durumun, sekizinci sınıf öğrencilerinin bu iki kümeye ayrıştırılmasında matematik başarısından çok, öğrencinin duyuşsal özelliklerinin katkısının olması ile bağlantılı olabileceği düşünülmektedir. Öğrencilerin duyuşsal özellikleri ile öğretmenin deneyim süresi ve öğrenim düzeyi arasındaki ilişkilere bakıldığında da çok düşük düzeyde ilişkilerin bulunduğu görülmektedir. Benzer şekilde, Gustafsson ve Nilsen (2016) tarafından TIMSS-2007 ve TIMSS-2011 verilerine dayalı olarak yapılan çalışmada da, 8. sınıf düzeyinde öğretmenin öğrenim düzeyinin öğrencilerin matematik başarısını açıklamaya az da olsa manidar düzeyde katkısının bulunduğu ancak deneyim süresinin matematik başarısı üzerinde etkisinin olmadığı gözlenmiştir. Bu bulgu da, sekizinci sınıf düzeyinde ortaya çıkan bu durumu destekler niteliktedir.

Bu araştırmada öğretimsel kalitenin göstergeleri olarak; anket maddelerinde belirtilen öğretimsel etkinlikleri gerçekleştirme sıklığı, öğretmenler arası etkileşim ve matematik dersinde ilgi çekici öğretime ilişkin görüşler ele alınmıştır. Öğretimsel etkinlikleri gerçekleştirme sıklığı ve öğretmenler arası etkileşim öğretmen yanıtlarına dayalı olarak belirlenirken, matematik dersinde ilgi çekici öğretime ilişkin görüşler öğrenci yanıtlarına dayalı olarak belirlenmiştir. Araştırmada her iki sınıf düzeyi için öğrencilerin sınıflanmasında matematik dersinde ilgi çekici öğretime ilişkin görüşlerin etkili olduğu ancak öğretimsel kalitenin diğer iki göstergesinin etkili olmadığı gözlenmiştir. Benzer şekilde Blömeke ve diğerleri (2016) tarafından yapılan çalışmada öğretmen kalitesinin göstergelerinin matematik başarısı ile ilişkili olduğu ancak öğretmen beyanına dayalı olarak belirlenen öğretimsel niteliklerin matematik başarısı ile ilişkili olmadığı gözlenmiştir. Bu doğrultuda öğrenci algısına dayalı olarak ele alındığında öğretimsel kalitenin, öğrencilerin sınıflanmasında önemli bir faktör olduğu ve bu üç kümedeki öğrencilerin matematik öğretiminin ilgi çekici olması konusundaki görüşleri açısından farklılaştıkları söylenebilir. Ancak öğretimsel etkinliklerin gerçekleştirilme sıklığı ve öğretmenler arası etkileşimin öğrencilerin sınıflamasında etkili olmaması, bu araştırmada ortaya çıkan üç kümedeki öğrenciler için bu niteliklerin farklılık sergilemediğine işaret etmektedir. Bu durumun, öğretmen niteliklerinde olduğu gibi, öğretmen beyanına dayalı belirlenen öğretimsel nitelikler ile öğrencilerin duyuşsal özellikleri arasında çok düşük düzeyde ilişkilerin bulunması ile bağlantılı olabileceği düşünülmektedir. Bu bulgular doğrultusunda dördüncü sınıf düzeyinde öğrenci profillerinin oluşturulmasında hem öğrenci düzeyindeki niteliklerin hem de sınıf düzeyinde öğretmen niteliklerinin etkili olduğu; sekizinci sınıf düzeyinde ise yalnızca öğrenci düzeyindeki niteliklerin etkili olduğu sonucuna ulaşılmıştır. Bu farklılığın ise öğrencilerin psiko-sosyal gelişim özellikleri ile ile ilişkili olabileceği düşünülmektedir.

Araştırmada öğrenci çıktıları, öğretmen nitelikleri ve öğretimsel nitelikler ile öğrenciye ilişkin duyuşsal özellikler açısından benzerlik ve farklılıklara dayalı olarak öğrenci profilleri oluşturulmaya çalışılmıştır. Tanımlanan öğrenci profilleri incelendiğinde her iki sınıf düzeyinde de matematikte en başarılı öğrencilerin, matematikte kendine çok az güvenen ancak matematik öğrenmeyi seven, okula aitlik hissi yüksek düzeyde olan ve akran baskısına düşük düzeyde maruz kalan öğrenciler olduğu görülmektedir. Matematikte başarılı olan öğrencilerin aynı zamanda matematik öğrenmeyi sevmesi, matematiğe değer vermesi, kendilerini okula ait hissetmesi ve akran baskısına çok az maruz kalmaları, ilgili öğrenci özellikleri ile matematik başarısı arasında ilişkilerin bulunduğuna işaret etmektedir. Bu bulgular, önceki araştırmalarda (Doğan ve Barış, 2010; Lee, 2009; Nortvedt ve

diğerleri., 2016; Sharkey, You ve Schnoebelen, 2008) söz konusu özellikler arasında gözlenen ilişkilerle tutarlılık sergilemektedir. Öğrencilerin okul ortamına ilişkin duyguları ve öğretmenleriyle ilişkileri onlar için önemlidir. Bu bağlamda öğretmenini ve okulunu seven öğrenciler, okula daha fazla aidiyet hissetmektedirler (Sharkey ve diğerleri., 2008). Okulunu, öğretmenini ve arkadaşlarını seven öğrencilerin eğitime ilişkin düşüncelerinin ve buna bağlı olarak sergiledikleri çabanın değişim gösterdiği ve bu durumun da matematik başarısı açısından da farklılıklara yol açmış olabileceği düşünülmektedir. Matematikte başarılı olan öğrenciler aynı zamanda akran baskısına en az maruz kalan öğrencilerdir. Akran baskısı ile öğrencilerin akademik başarısı arasındaki ters yönlü ilişki, ulusal ve uluslararası düzeyde gerçekleştirilen çalışmalarla (Nortvedt ve diğerleri., 2016; L. Rutkowski ve D.Rutkowski, 2016) da ortaya konulmuştur. Rutkowski ve Rutkowski (2016) yaptıkları çalışmanın sonuçlarına dayalı olarak akran zorbalığının öğretimsel süreçler aracılığıyla öğrenci başarısını etkilediğini belirtmişlerdir.

Her iki sınıf düzeyinde de öğrencilerin matematikte başarılı olmalarına rağmen bu alanda kendilerine güvenmemeleri ise ilgi çekici bir bulgudur. Lee (2009) tarafından, TIMSS-2011 uygulamasına dayalı olarak yapılan çalışmada da benzer bulgular elde edilmiştir. Araştırmacı Asya ülkelerinde öğrencilerin matematik yeterlik algısı ile matematik başarısı arasında ters yönlü ilişkiler gözlenmesine rağmen, Avrupa ülkelerinde bu iki özellik arasında pozitif yönde ilişkilerin gözlendiğini belirtmiştir. TIMSS-2015 uygulamasından elde edilen sonuçlara bakıldığında da matematikte kendine çok güvenen öğrenci yüzdesi açısından üst sıralarda yer alan Norveç, Hollanda, Bulgaristan, İngiltere gibi ülkelerde matematik başarı ortalamasının TIMSS matematik genel ortalamasından yüksek olduğu görülmektedir. Ancak kendine güven açısından üst sıralarda yer alan Kuveyt, Türkiye ve Jordan gibi Asya ülkelerinde ise matematik başarı ortalamasının TIMSS matematik genel ortalamasından düşük olduğu görülmektedir (Martin, Mullis, Foy ve Hooper, 2016). Bu örüntü, Asya ülkelerinde egemen olan doğu kültürünün özelliklerinin, başarı ve kendine güven arasındaki ilişki üzerinde etkilerinin olduğuna işaret etmektedir.

Sınıf düzeyindeki nitelikler açısından incelendiğinde, her iki sınıf düzeyi için de, bu gruptaki öğrencilerin öğretmenlerinin tamamının lisans düzeyinde öğrenim görmüş olduğu anlaşılmaktadır. Ancak matematikte başarılı olan öğrencilerin, öğretmenin deneyim süresi ve öğretimsel kalitenin öğretmen beyanına dayalı olarak belirlenen göstergeleri açısından sınıf düzeyleri (4. ve 8. sınıf) arası farklılıklar sergiledikleri gözlenmiştir. Dördüncü sınıf düzeyi için matematikte en başarılı olan öğrencilerin öğretmenleri ilgili öğretimsel etkinlikleri en fazla uygulayan, diğer öğretmenlerle etkileşim düzeyi en fazla olan ve ortalama 15 yıllık deneyime sahip öğretmenlerdir. Ancak sekizinci sınıf düzeyinde matematikte başarılı olan öğrencilerin öğretmenlerinin, öğretimsel etkinlikleri daha az sıklıkta uyguladığı, daha az deneyime (ortalama yaklaşık dokuz yıllık deneyim) sahip oldukları ve diğer öğretmenlerle daha az etkileşimde bulundukları gözlenmiştir. Öğretmenlerin öğretim sürecini planlamasında ve yönetiminde, okulun eğitim politikası ile öğretmen ve okulun beklentileri de önemli rol oynamaktadır. Ülke içerisinde eğitim düzeyi (ilkokul-ortaokul) açısından ve hatta okul düzeyinde bile bu bağlamda farklılıklar gözlenebilmektedir. Bu farklılıklar, araştırma kapsamında ele alınmayan ancak öğrenci çıktılarını etkileyebilecek değişkenlerin (örneğin, okul yönetiminin öğretmenlerin mesleki gelişim etkinliklerine katılımını desteklemesi), öğretimsel nitelikler ile öğrenci çıktıları arasındaki ilişkiler üzerindeki etkisini arttırarak, bu ilişkilerin sınıf düzeyine göre farklılaşmasına yol açmış olabilir (Blömeke ve diğerleri., 2016).

Bu araştırmada öğretimsel kalite, öğrenci yanıtlarına dayalı olarak da incelenmiştir. Bu bağlamda her iki sınıf düzeyi için de, matematikte başarılı öğrencilerin matematik derslerinde ilgi çekici öğretim yapıldığını düşündükleri gözlenmiştir. Bu bulgu aynı zamanda, öğrenci algısına göre öğretimsel kalitenin göstergesi ile matematik başarısı arasında, hem dördüncü hem de sekizinci sınıf düzeyi için, pozitif yönde ilişkinin varlığına işaret etmektedir. Bu bulguyu destekleyecek şekilde, Nortvedt, Gustafsson ve W.Lehre (2016) de TIMSS-2011 uygulamasına dayalı olarak yaptıkları araştırmada, öğrenci algısına göre belirlenen öğretimsel niteliğin, araştırmaya katılan ülkelerin %40'ında matematik başarısı ile pozitif yönde ilişkili olduğunu bulmuşlardır.

Matematikte başarı ortalaması en düşük olan öğrencilerin profilleri incelendiğinde ise, dördüncü ve sekizinci sınıf düzeyinde, bu öğrencilerin duyuşsal özellikler açısından benzer nitelikte oldukları görülmektedir. Her iki sınıf düzeyi için de bu gruptaki öğrencilerin matematikte kendine çok fazla güvendikleri ancak matematik öğrenmeyi sevmedikleri, kendilerini öğrenim gördükleri okula ait hissetmedikleri, matematiğe değer vermedikleri ve akran baskısına yüksek düzeyde maruz kaldıkları anlaşılmaktadır. Bu bulgular, alanyazınla (Doğan ve Barış, 2010; Lee, 2009; Nortvedt ve diğerleri., 2016; Sharkey ve diğerleri., 2008) tutarlı olarak, öğrenci düzeyindeki bu nitelikler ile öğrencinin matematik başarısı arasında ilişkilerin bulunduğuna işaret etmektedir. Bu gruptaki öğrenciler için, öğretmen nitelikleri ve öğretimsel nitelikler açısından bir inceleme yapıldığında ise sınıf düzeyine göre farklılıkların olduğu görülmektedir. Dördüncü sınıf düzeyinde matematikte düşük başarı sergileyen öğrencilerin öğretmenlerinin çok büyük bir kısmının lisans düzeyinde öğrenim görmüş ve ortalama 11 yıllık deneyime sahip oldukları görülmektedir. Sekizinci sınıf düzeyinde ise bu gruptaki öğrencilerin öğretmenlerinin büyük bir kısmı lisans, çok az bir kısmı önlisans (Yüksekokul-2 yıllık) ve yine çok az bir kısmı lisansüstü düzeyde öğrenim görmüş kişilerdir. Bu öğretmenler de ortalama 11 yıllık deneyime sahip kişilerdir. Öğretimsel kalite açısından bakıldığında dördüncü sınıf düzeyinde bu öğretmenlerin, ilgili öğretimsel etkinlikleri daha az sıklıkta gerçekleştirdikleri ve diğer öğretmenlerle etkileşim düzeylerinin de düşük olduğu anlaşılmaktadır. Sekizinci sınıf düzeyinde ise, matematikte düşük başarı gösteren öğrencilerin öğretmenlerinin, ilgili öğretimsel etkinlikleri biraz daha sık gerçekleştirdikleri ve diğer öğretmenlerle daha fazla etkileşimde oldukları bulunmuştur. Öğrenci görüşlerine göre öğretimsel kalite açısından bakıldığında ise her iki sınıf düzeyinde de matematikte başarı düzeyi düşük öğrencilerin matematik dersindeki öğretimin ilgi çekici olmadığını düşündükleri görülmektedir.

Araştırma bulguları, araştırmada esas alınan kavramsal çerçevedeki ilişkileri destekleyecek şekilde, dördüncü sınıf öğrencileri için öğrenci çıktılarındaki bireysel farklılıkların sınıf düzeyinde öğretmen nitelikleri ve öğrenci düzeyinde ise öğrencinin duyuşsal özellikleri ile bağlantılı olduğunu göstermektedir. Ancak sekizinci sınıf öğrencileri için elde edilen bulgular, öğrenci çıktılarının öğretmen kalitesinin göstergelerinden ziyade öğrenci düzeyinde ele alınan duyuşsal özellikler ile daha fazla ilişkili olduğunu ortaya koymaktadır. Ancak araştırma sonuçları her iki sınıf düzeyi için de öğretimsel kalitenin öğretmen beyanına dayalı olarak incelenen göstergelerinin, öğrencilerin öğrenme çıktıları (matematik başarısı) ile düşük düzeyde ilişkili olduğuna işaret etmektedir. Benzer şekilde Blömeke ve diğerleri (2016) de öğretmen beyanına dayalı olarak belirlenen öğretimsel kalitenin, öğrenci öğrenme çıktıları ile manidar düzeyde ilişkili olmadığını ve bu durumun TIMSS öğretmen anketindeki öğretimsel etkinliklere ilişkin maddelerin, öğretimsel kaliteyi tüm yönleri ile ele alıp ölçememesinden kaynaklanmış olabileceğini rapor etmişlerdir. Bu çalışmada da öğretim kalitesinin dolaylı ölçümleri ele alınmıştır. Öğretim kalitesinin öğretmen ve öğrenci algısına dayalı olarak dolaylı bir şekilde ölçülmesi yerine, kalite göstergelerinin doğrudan ölçümleri elde edilerek öğrenci çıktıları ile ilişkileri incelenebilir. Bunun yanısıra alanyazınla (Nortvedt ve diğerleri., 2016) tutarlı olarak bu çalışmada, öğretimsel kalitenin öğrenci görüşlerine dayalı olarak belirlenen göstergesinin öğrenci öğrenme çıktıları ile ilişkili olduğu sonucuna ulaşılmıştır.

Araştırmada her iki sınıf düzeyi için de öğrenci çıktılarındaki farklılıkların, öğrencilerin duyuşsal özellikleri ve matematik öğretiminin niteliğine ilişkin görüşlerindeki farklılıklar ile birlikte gözlendiğini ortaya koymaktadır. Bu sonuçlar, öğrenci çıktılarının en önemli belirleyicilerinin öğrencilerin duyuşsal özellikleri ile matematik öğretiminin niteliğine ilişkin görüşlerinin olduğuna işaret etmektedir. Bu doğrultuda öğretim sürecinin planlanmasında öğretmenlerin matematik eğitiminin duyuşsal boyutunu dikkate alarak, öğrencilerin matematik alanına ilgi duymasını sağlayacak ve matematik öğrenmelerini teşvik edecek çeşitli etkinliklere yer vermeleri önerilebilir. Okul yöneticileri ve öğretmenler, öğrencilerin okulu daha fazla benimsemelerini, akranlarla olan ilişkilerini geliştirmelerini sağlayacak ve okul ortamının öğrenciler için daha keyif alacakları ortamlar olmasını sağlayacak etkinlikler düzenleyebilirler. Bu etkinlikler, hem öğrenciler arası olumlu ilişkilerin gelişmesine katkıda bulunarak akran baskılarının azalmasını hem de öğrencilerin okul aidiyeti hislerinin güçlenmesine katkı sağlayacaktır. Akran baskısının ve okula aidiyet hissinin matematik başarısı ile ilişkileri düşünüldüğünde bu etkinliklerin, öğrenci öğrenme çıktılarına katkı sağlayacabileceği düşünülmektedir. Araştırmada ayrıca sınıf düzeyine göre öğretmen niteliklerinin,

öğrenci öğrenme çıktılarını etkileme biçiminin farklılaştığı ve dördüncü sınıf düzeyi için öğretmenin eğitimsel geçmişinin belirleyici nitelikte olmasına rağmen sekizinci sınıf için bu durumun geçerli olmadığı gözlenmiştir. İlgili alanyazının da bu sonuçları desteklemesine dayalı olarak araştırmacılara sınıf düzeyi için bu farklılaşmanın olası sebeplerini ayrıntılı bir şekilde incelemek üzere görüşme yöntemi, odak grup yöntemi gibi yöntemlerin kullanılması önerilebilir. Bu araştırmada öğrenci çıktılarını etkileyebilecek sınıf düzeyindeki nitelikler ile öğrenci düzeyindeki nitelikler açısından öğrenci profilleri oluşturulmaya çalışılmıştır. İleride yapılacak araştırmalarda, ulusal düzeydeki nitelikler ile okul düzeyindeki nitelikler ve öğrenci çıktıları açısından benzerliklere ve farklılıklara dayalı olarak öğrenci profillerinin oluşturulması önerilebilir.

Araştırmada ele alınan kavramsal çerçevede öğretmenin eğitimsel geçmişi, öğretimsel süreçler için kaynak olarak düşünülmektedir. Bu bağlamda öğretmenin alan bilgisinin ve eğitim alanındaki bilgi ve becerilerinin öğretim sürecine yansımasını inceleme imkanı sağlayan TIMSS gibi geniş ölçekli testler, araştırmacılar için önemli veri kaynağı niteliğindedir. Bu doğrultuda bu tür geniş ölçekli test uygulamaları öğretmen kalitesinin, hem öğretimsel kaliteye hem de öğrenci çıktılarına etkilerinin ulusal ve uluslarası düzeyde incelenmesine imkan sağlamaktadır. Bu çalışmaların sonuçlarının, öğretmen eğitiminin niteliğine ilişkin de bilgi verdiği ve öğretmen eğitimi programlarının geliştirilmesine de katkı sağladığı düşünülmektedir. Hali hazırdaki çalışmanın sonuçları ise, öğretmenin eğitimsel geçmişinin ve öğrencilerin duyuşsal özellikleri ile matematik öğretimine ilişkin görüşlerinin eğitsel çıktılar açısından önemini ortaya koymakta ve TIMSS gibi geniş-ölçekli testlerin bu bağlamdaki katkılarını göstermektedir. Geniş ölçekli testler aracılığıyla eğitim politikaları, araştırmaları ve uygulamalarına ilişkin önemli sorunların ve soruların çalışılması sağlanmaktadır.

KAYNAKÇA

Akyüz, G. (2006). Investigation of the effect of teacher and class characteristics on mathematics achievement in Turkey and European Union Countries. *Elementary Education Online, 5*(2), 75-86.

Berberoğlu, G., Çelebi, Ö., Özdemir, E., Uysal, E., & Yayan, B. (2003). Üçüncü Uluslararası Matematik ve Fen çalışmasında Türk öğrencilerinin başarı düzeylerini etkileyen etmenler. *Eğitim Bilimleri ve Uygulama, 2*(3), 3-14.

Blömeke, S., Olsen, R.V. and Suhl, U. (2016). Relation of student achievement to the quality of their teachers and instructional quality. In T. Nilsen, J.E. Gustafsson (Ed), *Teacher quality, ınstructional quality and studentt outcomes,* (Vol. 2, 21-50). Switzerland: Springer International Publishing.

Bos, K., & Kuiper, W. (1999). Modeling TIMSS data in a European comparative perspective: exploring influencing factors on achievement in mathematics in grade 8. *Educational Research and Evaluation, 5(2),* 57-179. doi: 10.1076/edre.5.2.157.6946

Buluç, B. (2014). TIMSS 2011 sonuçları çerçevesinde okul iklimi değişkenine göre öğrencilerin matematik başarı puanlarının analizi. *Gazi Üniversitesi Endüstriyel Sanatlar Eğitim Fakültesi Dergisi, 33,* 105-121.

Butakor, P. K. (2016). Hierarchical linear modeling of the relationship between attitudinal and instructional variables and mathematics achievement. *International Journal of Research in Education Methodology, 7(5),* 1328-1336. doi:10.24297/ijrem.v7i5.4342

Büyüköztürk, Ş. (2012). *Veri analizi el kitabı*. Ankara: Pegem.

Creemers, B. and Kyriakides, L. (2009). Situational effects of the school factors included in the dynamic model of educational effectiveness. *South African Journal of Education, 29,* 293-315.

Eklöf, H. (2007). Self concept and valuing of mathematics in TIMSS 2003: Scale structure and relation to performance in a Swedish setting. *Scandinavian Journal of Educational Research*, *51*(3), 297–313. doi:10.1080/00313830701356141

Ercikan, K., Simon, M. And Oliveri, M.E. (2013). Score compability of multiple language versions of assessments within jurisdictions. In M. Simon, K.Ercikan and M.Rousseau (Ed), *Improving large-scale assessment in education,* (1st ed., 110-124). New York: Taylor&Francis.

Erikson, H. E. (1968). *Identity: Youth and crisis*. New York: W. W. Norton.

Garson, G. D. (2015). *Missing value analysis and data ımputation*. Ashebora: Statistical.

Gustafsson, J.E. and Nilsen, T. (2016). The impact of school climate and teacher quality on mathematics achievement: A difference-in-difference approach. In T. Nilsen, J.E. Gustafsson (Ed), *Teacher quality,*

instructional quality and student outcomes, (Vol.2, 81-95). Switzerland: Springer International Publishing.

Hong, H.K. (2012). Trends in Mathematics and Science Performance in 18 Countries: Multiple regression analysis of the cohort effects of TIMSS 1995-2007. *Education Policy Analysis Archives, 20*(33). doi: 10.14507/epaa.v20n33.2012

House J. D. & Telese, J. A. (2008). Relationships between student and instructional factors and algebra achievement of students in the United States and Japan: An analysis of TIMSS 2003 data, educational research and evaluation. *An International Journal on Theory and Practice*, *14*(1), 101-112. doi: 10.1080/13803610801896679

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, andmath anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences, 19,* 355–365. doi: 10.1016/j.lindif.2008.10.009

Little, J. W. (1981). *The power of organisational setting, school success and staff development*, Washington DC: National Instıtude of Education.

Martin, M. O., Mullis, I. V. S. and Hooper, M. (2016). (Eds.), *Methods and procedures in TIMSS 2015.* Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Martin, M.O, Mullis, I.V.S., Foy, P. ve Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/wp-content/uploads/filebase/full%20pdfs/T15-International-Results-in-Mathematics.pdf.

Nilsen, T., Gustafsson, J.E., Blömeke, S. (2016). Conceptual framework and methodology of this report. In T. Nilsen, J.E. Gustafsson (Ed), *Teacher quality, ınstructional quality and studentt outcomes,* (Vol.2, 1-19). Switzerland: Springer International Publishing.

Nortvedt, G.A., Gustafsson, J.E. and Lehre, W. (2016). The importance of instructional quality for the relation between achievement in reading and mathematics. In T. Nilsen, J.E. Gustafsson (Ed), *Teacher quality, ınstructional quality and studentt outcomes,* (Vol.2, 97-113). Switzerland: Springer International Publishing.

Özdamar, K. (2004). Paket *programlar ile istatistiksel veri analizi 2*. Eskişehir: Kaan Kitabevi.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. doi: 10.1016/0377-0427(87)90125-7

Rutkowski, L. and Rutkowski, D. (2016). The relation between studets' perceptions of instructional quality and bullying victimization. In T. Nilsen, J.E. Gustafsson (Ed), *Teacher quality, ınstructional quality and studentt outcomes*, (Volume 2, 115-133). Switzerland: Springer International Publishing.

Sharkey, J. D., You, S. ve Schnoebelen, K. (2008). Relations among school assets, individual resilience, and student engagement for youth grouped by level of -family functioning. *Psychology in the Schools*, *45*(5), 402-418. doi:10.1002/pits.20305

Shih, M-Y., Jheng, J-W & Lai, L-F. (2010). A two-step method for clustering mixed categroicaland numeric data. *Tamkang Journal of Science and Engineering*, *13*(1), 11-19.

Skinner, E., Furrer, C., Marchand, G. ve Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology, 100*(4), 765-781. doi:10.1037/a0012840

Stemler, E. S. (2001). *Examining school effectiveness at the fourth grade: A hierarchical analysis of the Third International Mathematics and Science Study (TIMSS).* (Doctoral dissertation, Boston College the Graduate School of Education). Retrieved from https://www.researchgate.net/publication/.

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region*. Melbourne: ACER and Bangkok, UNESCO. Retriewed from https://research.acer.edu.au/cgi/viewcontent.cgi?article=1020& context=monitoring_learning

Yalçın, S., Demirtaşlı, R.N., Dibek, M.I. ve Yavuz, H.Ç. (2017). The effect of teacher and student characteristics on tımss 2011 mathematics achievement of fourth-and eighth-grade students in Turkey *International Journal of Progressive Education*, *13*(3), 79-94.

Yayan, B. (2003). *A cross-cultural comparison of mathematics achievement in the Third International Mathematics and Science Study-Repeat (TIMSS-R) (*Yüksek Lisans Tezi, Orta Doğu Teknik Üniversitesi, Ankara). Erişim adresi: http://tez.yok.gov.tr/

Milli Eğitim Bakanlığı. (2016). *TIMSS-2015 ulusal matematik ve fen ön raporu*. http://timss.meb.gov.tr/wp-content/uploads/TIMSS_2015_Ulusal_Rapor.pdf adresinden erişildi.

Yoshino, A. (2012). The relationship between self-concept and achievement in TIMSS 2007: A comparison between American and Japanese students. *Inernational Review of Education, 58,* 199-219. doi: 10.1007/ s11159-012-9283-7

EXTENDED ABSTRACT

Introduction

Large-scale tests have been used for making decisions related to educational policies, accountability and improving educational systems for the last three decades. Based on these tests' results, countries could determine the deficiencies and problems of their educational systems and take initiatives towards these problems (Ercikan, Simon, and Oliveri, 2013). The best known large-scale tests are PISA, PIRLS, TALIS and TIMSS. TIMSS provides information related to student achievement, student attributions, teacher qualifications, and school qualifications.

Fundamentally, TIMSS is developed and applied to assess students' acquired knowledge and skills in mathematics and science. As TIMSS provides information related to student outcomes, teacher qualifications, instructional qualifications and school qualifications at national and international levels, it gives researchers opportunities to investigate the relations among these qualifications (Nilsen, Gustaffson and Blömeke, 2016). The relations among these factors have also been shown by the studies (Akyüz, 2006; Berberoğlu, Çelebi, Özdemir, Uysal and Yayan, 2003; Blömeke, Olsen and Suhl, 2016; Bos and Kuiper, 1999; Gustafsson and Nilsen, 2016; Houseand Telese, 2008; Rutkowski and Rutkowski, 2016) conducted on this topic. In the literature there are theoretical models developed to explain how these factors influence student outcomes. Of those models the most comprehensive one is "the Dynamic Model of Educational Effectiveness" which has been developed by Creemers and Kyriakides' (2006) (quoted from Creemers and Kyriakides, 2009). Based on this model, Nilsen and others (2016) have formed a conceptual framework for the determinants of student outcomes. In this conceptual framework, the determinants of the student outcomes are examined at national-level, school-level, class-level and student-level and it is hypothesized that the qualifications at these levels are inter-related and also related to the student outcomes. At national-level, the differences related to cultural factors, educational values, educational policies and educational systems are considered. At school-level the factors regarding school's academic and social climate; at class-level teacher qualifications and instructional qualifications and at student-level students' attributions are taken in to account. Academic achievement and perceived peer pressure are examined as student outcomes (Nilsen and others., 2016).

Based on this conceptual framework, in this study, the math achievement and perceived peer pressure were examined as student outcomes while the student affective attributions, teacher qualifications and instructional qualifications were examined as the determinants of student outcomes. Accordingly, it was aimed to create student profiles by clustering students based on the student outcomes and student and teacher qualifications and instructional qualification that could influence these outcomes. It was considered that investigating the similarities and differences among students, in terms of mathematics achievement and the class-level and student-level qualifications that might influence math.achievement it, is important to enhance student educational outcomes. Additionally, based on the findings of such studies, educators and policy makers may take various initiatives in order to enhance the quality of mathematics and science education.

Method

Study Group

In this study, initially, an analysis was carried out for the missing values. Afterwards, the teachers for whom there were missing values for the variables of "education level" and "job experience" and the students of these teachers were not included in the study. Accordingly, the study group was

comprised of n=244 teachers and n=6267 students from the 4th grade level and n= 216 teachers and n= 5965 students from the 8th grade level.

Procedure and Data Analysis

In this study, data were obtained from the mathematics achievement test used in TIMSS-2015 and the scales in the Teacher Questionnaire and Student Questionnaire. The data related to teacher education and teacher job experience were directly obtained from TIMSS-2015 international database. For other variables, index scores (z scores) were produced via Exploratory Factor Analysis (EFA) based on the responses given to items in the related scales in Teacher and Student Questionnaires. As indicators of mathematics achievement, Plausible Values obtained from Math Achievement Test were taken. A series of EFA's results revealed that these scales and Math Achievement Test could yield valid measures of related constructs. Although some Cronbach α values calculated for these measures were low, in general it could be concluded that these scales could give reliable (at least acceptable level) measures of the related constructs (α=0.53- α=0.98). Afterwards, for classifying students in terms of the mathematics achievement, student's affective attributions, teacher qualifications and instructional qualifications, two-step clustering method was used.

Results and Discussion

In this study, the 4th grade and 8th grade students in the study group were clustered in terms of student outcomes, teacher qualifications and instructional qualifications and student profiles were created. The defined profiles for the students at fourth and eight grade levels are similar in terms of the student mathematics achievement and student attributions. At both grade levels, students with high-level achievement are those who have low-confident in mathematics, like learning mathematics, have a sense of belonging to the school and who have slightly been exposed to peer pressure. This finding is consistent with both the conceptual framework examined in this study and the related literature. On the other hand, it has been found that teacher qualifications of these students differ by grade level. At fourth-grade level, all teachers have bachelor's degree and they have 15-years job experience on average. They are the ones who use the instructional practices most frequently and who interact with other teachers most among the fourth grade-level teachers in the study group. At the eight-grade level, almost all of the teachers of the students in this class have bachelor's degree and they have 9-years job experiences on average. These teachers have reported that they have used the instructional practices less frequently and they have interacted with other teachers randomly.

Study findings showed that the individual differences in the student outcomes of fourth-grade students were related to the teacher qualifications and students' affective attributions, which supported the relations in the conceptual framework. However, the findings for eight-grade level students demonstrated that student outcomes were much more related to students' affective attributions rather than teacher qualifications. Additionally, findings indicated that there were weak relationships between the indicators of instructional quality (based on teachers' report) and student learning outcomes. According to the results of this study, it could be suggested to the teachers to design and perform various activities that increase students' interest in mathematics, to encourage them to learn mathematics and help them enjoy learning mathematics while organizing instructional process.